THE COMPARATIVE EFFECTIVENESS OF DIFFERENT ITEM ANALYSIS TECHNIQUES IN INCREASING CHANGE SCORE RELIABILITY

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
LINDA D. MITCHELL
1970

This is to certify that the

thesis entitled

THE COMPARATIVE EFFECTIVENESS OF DIFFERENT ITEM ANALYSIS TECHNIQUES IN INCREASING CHANGE SCORE RELIABILITY

presented by

Linda D. Mitchell

has been accepted towards fulfillment of the requirements for

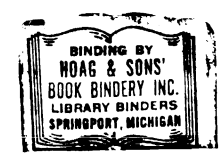
Ph.D. **degree in** Education

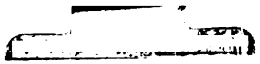
William McBrewster

Major professor

Date July 1, 1970

O-169





ABSTRACT

THE COMPARATIVE EFFECTIVENESS OF DIFFERENT ITEM ANALYSIS TECHNIQUES IN INCREASING CHANGE SCORE RELIABILITY

By

Linda D. Mitchell

Four different procedures for selecting items to measure individual change were studied to determine which would result in sets of items with the highest change score reliability. The four methods of item analysis used for these change items were: selection on the basis of change item score variance; selection on the basis of pretest response frequency; selection on Saupe's correlation between change item score and total score; and selection on triserial correlation.

The study was specifically undertaken to determine whether these methods of change item analysis could lead to the selection of more reliable subsets of items than could be obtained by randomly choosing items from a pool. Comparisons between the different methods were also made. The sample used for item analysis and

cross-v

Universi

men in 1

group. C

conducte

each proc

control pr

subsets.

one dichot

two differ

T

scored for

were calcu

descriptiv

the entire

hypotheses

smaller in

subsets ch

Hypotheses

Tukey post

cross-validation was a group of 263 students at Michigan State University who had been tested on the Inventory of Beliefs as freshmen in 1958, and again as juniors in 1961.

Half of this sample were assigned to an initial item analysis group. On the basis of their responses the four item analyses were conducted and subsets of 15, 30, 60, and 90 items were chosen by each procedure from the original pool of 120 items. In addition, a control procedure of random selection was also used to select item subsets. Items were scored on both a one-to-four scale and a zero-one dichotomy. Item analyses were carried out separately for these two different scoring procedures.

The items selected by the item analysis methods were then scored for the cross-validation group. Change score reliabilities were calculated based upon these responses. To obtain the best descriptive comparison, all reliability estimates were computed for the entire cross-validation group of 131 students. To test the hypotheses of the study, the cross-validation group was divided into smaller independent samples and change score reliabilities for item subsets chosen by different methods were computed on these samples. Hypotheses were tested by using a two-way analysis of variance with Tukey post hoc comparisons for mean differences.

scored o

resulted

random s

high chan

quency ar

methods o

than did r

ance, sele

this case,

No signifi

methods of

The results of the analysis showed that when the items were scored on a one-to-four scale, three methods of item analysis resulted in significantly higher change score reliability than did random selection. Saupe's r_{dD} was the most successful in producing high change score reliability. Selection on the basis of pretest frequency and change score variance were equally effective.

When the items were scored on a zero-one basis, three methods of item analysis resulted in greater change score reliability than did random selection. These were: selection on change variance, selection on pretest frequency, and triserial correlation. In this case, Saupe's correlation was not superior to random selection. No significant differences were found between the three successful methods of change item analysis.

TH

THE COMPARATIVE EFFECTIVENESS OF DIFFERENT ITEM ANALYSIS TECHNIQUES IN INCREASING CHANGE SCORE RELIABILITY

By

Linda D. Mitchell

A THESIS

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

DOCTOR OF PHILOSOPHY

Department of Educational Psychology

1970

Mehrens,
assistanc
work and
suggestio
Committe
Bell--ar
to Dr. Ir
data used
enabled
and rese

ACKNOWLEDGMENTS

The author expresses her sincere thanks to Dr. William A. Mehrens, Chairman of the Guidance Committee, for his counsel and assistance throughout her doctoral program and in the experimental work and preparation of the manuscript for this study. The helpful suggestions and editorial comments of members of the Guidance Committee--Dr. Andrew Porter, Dr. Leroy Olson, and Dr. Norman Bell--are gratefully acknowledged. Special thanks is also extended to Dr. Irvin J. Lehmann, who generously provided access to the data used in this study.

The financial support of an NDEA Title IV Fellowship enabled the author to carry out her doctoral program of coursework and research at Michigan State University.

LIST OF

CHAPTE

I.

II.

III.

IV.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

TABLE Page

research

change f

this prob

individua

where D

score at

score.

however

These c

ment ex

recogni

primar

CHAPTER I

THE PROBLEM

A methodological problem frequently encountered by researchers in education is how to obtain measures of growth or change for subjects over a given period of time. One approach to this problem has been to calculate the change score for each individual, using the formula:

$$D = X - Y , \qquad (1)$$

where D is the change score, Y is the score at time 1, and X is the score at time 2. D has also been called a gain score, or discrepancy score.

Researchers who have attempted to use such change scores, however, have been plagued by one persistent psychometric problem. These change scores are remarkably unreliable. Noted measurement experts such as Lord, Horst, Webster, and Bereiter have long recognized this problem (Harris, 1963). When the researcher is primarily interested in measuring change for a group, this problem

of low reliability is not too serious; however, if he wishes to make meaningful comparisons between individuals on the basis of their growth or attitude change, then the lack of reliability becomes crucial.

When the traditional formula for the reliability of change scores is examined, two factors seem to be necessary to obtain high change score reliability. This formula as derived by Gulliksen (1950, p. 353) is:

$$r_{DD} = \frac{\bar{r} - r_{XY}}{1 - r_{XY}} , \qquad (2)$$

where $\bar{r}$ is the mean of r_{XX} and r_{YY}. From this it appears that in order to obtain a high value for r_{DD}, the internal consistency of the test at time 1 (r_{YY}) and at time 2 (r_{XX}) should be high, but the stability coefficient for the test over time (r_{XY}) should be somewhat lower. Thus change score reliability can be increased if the test-retest correlation can be reduced while test homogeneity (or internal consistency) is maintained at a high level for each separate administration of the test.

When the reliability of an instrument is unsatisfactory, a common psychometric practice is to construct more items, since longer tests are usually more reliable. The obvious drawback in

this procedure is that for many testing situations the number of items must be kept to a minimum for practical considerations of time and economy. When this is the case, item analysis techniques are usually employed to select subsets of the most discriminating items from the original pool so that the test can be shortened without seriously reducing its reliability.

Ordinary item analysis procedures, usually based upon a single test administration, are designed to improve test internal consistency or to yield a test which correlates highly with some criterion. Such methods are not guaranteed to work for change score reliability. Theorists such as Bereiter (1963), Saupe (1966 and 1961), and Lord (1968, p. 331) have suggested that a researcher who desires to construct an instrument, sensitive to individual change, should use item analysis techniques suited for that purpose.

Several new techniques for such item analyses have recently appeared in the literature. One of these methods is based upon observing the response changes to items over time (Gruber and Weitman, 1962). Items for which there is a "moderate" change frequency when a group is tested and retested at a later date should be selected. This tends to eliminate those items for which the group exhibited little change in response over time as well as items for which the group displayed a universal change over time. In other

words, items for which there is "moderate" rate of change will be items for which there was variation between subjects in their response changes.

A second method uses response frequency to items on the pretest only. With this method the expected direction of change must be known in advance (Gruber and Weitman, 1962). The experimenter then selects items which had a low percentage of negative responses on the pretest if a high percentage of negative responses is expected on the posttest, or vice versa.

In a third method items are selected which have a high correlation between item response change and total change score. This correlation is determined from a formula derived by Saupe (1966), which is equivalent to the Pearson Product Moment correlation value.

A fourth method of item analysis was employed in this study which had not been revealed when literature in this area was reviewed. With this method items are selected if they have high triserial correlation values when the correlation between total change score and trichotomized change in item response is computed.

Because of the relative newness of these item analysis methods there has been little empirical research to determine whether or not they could effect increases in change score reliability. Also, the comparative efficiency and effectiveness of these different

procedures is completely unknown. Such information is sorely needed by researchers who face the problem of constructing instruments to reliably measure growth or attitude change for individuals over time (Lord, 1968, p. 331).

To provide further information on this topic, an empirical study was designed to examine these various item analysis procedures and their effects upon change score reliability.

Purpose of This Study

The purpose of this study was to determine whether use of the item analysis methods previously discussed could increase the reliability of change scores on a collegiate attitude survey. Four specific questions of central importance to this issue were raised.

1. Which of these four item analysis methods would result in selecting a subset of items with the highest change score reliability?
2. Which correlational method would result in the higher estimate of change score reliability for selected subsets of items?
3. Would the response frequency method, based upon the variances of response changes from pretest to posttest, result in higher change score reliability than the method which uses only pretest response frequency?

4. Could reliability of change scores for items selected on the basis of pretest response frequency exceed the reliability of an equal number of randomly chosen items?

This fourth question was particularly interesting because of its practical significance for test construction. In many attempts to measure change the experimenter simply does not have time to construct his instrument and run a complete item analysis on test-retest data before he can gather his data. (This is especially true for longitudinal studies.) Thus, if a method could be developed to eliminate useless items on the basis of pretest characteristics alone, it would be extremely helpful and time-saving for the researcher and his subjects.

Hypotheses

On the basis of reliability and item analysis theory, four general hypotheses were formulated in an attempt to answer the questions under investigation in this study. These hypotheses were:

1. Use of Method III (computing the PPM correlation for total change score and item change response) would result in a subset of items with higher change score reliability than that of item subsets chosen by any other method or by random selection.

2. Method IV (computing the triserial correlation between change scores and change in item response) would result in a subset of items with higher change score reliability than could be obtained for items chosen by response frequency methods or by random selection.
3. Method I (selecting items which showed variance in changes in response frequency over time) would result in a subset of items with higher change score reliability than that of items selected by Method II (selection on the basis of pre-test response only).
4. A subset of items could be selected by Method II (pretest response frequency) which would have higher change score reliability than a randomly selected subset of items.

Theoretical Rationale

The idea of attacking the unreliability of change scores at the item level can be credited to Bereiter, who formulated the concept of the change item. The change item was defined in this way:

> A single item administered on two occasions yields an item change score which is the difference between the item scores on the two occasions. If the item is scored dichotomously, 1 or 0, on each occasion, then the change item may take any of three values, 1, 0, or -1. (Bereiter, 1963, p. 10)

This definition can be expanded to include items which have more than 0 or 1 as a possible score on each occasion, such as those found on many attitude scales. Change items may thus be scored for both direction and amount of change, and change item scores may be summed, like ordinary item scores, to get a total change score,

$$D = \sum d_i \tag{3}$$

where

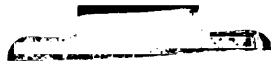
$$d_i = x_i - y_i . \tag{4}$$

In this definition y_i is the individual's response to item i at time 1, and x_i is his response to item i at time 2.

Bereiter believed that item analysis procedures could be carried out on the change items to improve change score reliability. Furthermore, he maintained that change score reliability could be adequately defined using a classical definition of reliability. Using change item scores the formula becomes:

$$r_{DD} = 1 - \frac{\sum S^2_{d_i}}{\sum S^2_{d_i} + \sum_{i \neq j} C_{d_i d_j}} \tag{5}$$

where $S^2_{d_i}$ is the variance of a change item and $C_{d_i d_j}$ is the covariance for the change item scores of items i and j. Bereiter then



hypothesized that increases in change score reliability could be attained by selecting items in such a way as to maximize the change item covariances.

Clearly the two methods of item analysis which use correlations between total change score and item change score as the indices for selecting items are directly based on this line of thought. Change items which have high correlations with total change score must have high intercorrelations with each other.

Further consideration of the two correlational indices for change item analysis reveals that they correspond directly to two popular indices often used for selection of regular dichotomous items. The familiar point-biserial correlation coefficient for dichotomous items is actually a Pearson Product Moment correlation (Magnusson, 1966, p. 199). In addition, the triserial correlation is derived using the same assumptions as the well-known biserial correlation, and the formulae for these two statistics are identical, except for the inclusion of the parameters for the third category in the triserial expression. Expressions for both biserial and triserial correlations can be derived from the general expression for the multiserial correlation coefficient given by Jaspen (1946). (A more complete discussion of this topic follows in the literature review in Chapter II.) These similarities should help to answer the question:

Which correlational method of change item analysis will result in greater change score reliability? Lord (1968, p. 344) pointed out that when there are ability differences between item analysis and cross-validation groups, the biserial correlation, which is unaffected by the factor of item difficulty, might be better for selecting items with high reliability across groups; however, when the groups are similar, the point biserial method might produce a more reliable test. In this experiment subjects from the same population were randomly assigned to item analysis and cross-validation groups. Since the two groups could be expected to be fairly similar, it was hypothesized that the PPM method (the point biserial method) would result in more reliable change scores than would the triserial index.

The rationale for selecting items on the basis of response frequency is apparent from Formula (5). One way to increase reliability is to increase the item covariance/variance ratio. Assuming item intercorrelations remain constant, this can be accomplished by selecting items which have large variances. As individual item variances are increased, item covariances must also increase, but the total of the item covariances will increase at a faster rate than the total of the item variance. The change items with the greatest variances will be those with moderate "difficulty" levels or frequencies of response change. This is clearly illustrated

if two extreme cases for change items are considered. An item for which there was no change in response between testings will have a mean change score of 0 and a change variance of 0. Likewise an item for which there was a universal shift for the group from positive to negative response will have a mean change score of 1 and a change variance of 0. Such items can contribute nothing to reliability (Shoemaker, 1969); however, items which have moderate frequencies of response changes will have variances which are larger and can reflect differences in individual changes which are necessary for high change item covariances and, consequently, high change score reliability.

If the direction of change cannot be predicted, it is impossible to choose items on the basis of pretest response frequency to insure that they will have adequate response change. If the direction of the change can be predicted, then the items can be chosen which are likely to have response shifts that will produce the desired rate of change. For example, if a shift toward a positive "Agree" response is expected over time, then items which initially have a high proportion of "Disagree" responses will be likely to have a moderate frequency of response changes over time. (Obviously the researcher must hope that a total shift to the "Agree" response does not occur.) This procedure is more risky than selecting items when

the actual response changes and their variance can be computed from a complete set of test-retest data.

It was generally expected that correlational methods would be superior to response frequency methods for item selection because the correlational procedures depend upon both item variances and their intercorrelations, while the frequency methods fail to consider how an individual item covaries with others in the item pool.

An Overview

Further discussion of theoretical works and empirical research studies which are related to the problem of selecting items to measure change and the reliability of change scores is presented in the Review of Literature, Chapter II. An empirical study designed to compare several different change item analysis methods is described in Chapter III. In Chapter IV the method of statistical analysis used to test the hypotheses of this study and the results of that analysis are presented. The conclusions from this study, discussion of the results, and some implications for future research in this area have been summarized in the fifth and final chapter.

CHAPTER II

REVIEW OF LITERATURE

Whenever the problem of measuring change is considered, the researcher must be careful to specify whether he wishes to evaluate mean change for a group or to study relative changes between individuals within a group. The need for this distinction has been pointed out by Lord (1963), Webster (1963), and Tucker et al. (1966). If individual differences are the main interest, then the researcher must be concerned with the reliability of his observations of change (Webster, 1963).

Traditionally difference scores have been regarded as so unreliable that Gulliksen (1950, p. 354) urged that standardized test publishers should warn their users of this fact and actually report difference score reliability in their technical manuals. Lord (1958) urged that counselors should make very cautious interpretations when advising individuals on the basis of difference scores.

Concern over the reliability of difference scores led to the development of several different expressions for its estimation.

One well-known expression for this reliability was given by Gulliksen (1950, p. 353):

$$r_{DD} = \frac{\bar{r} - r_{XY}}{1 - r_{XY}} . \qquad (6)$$

The value for $\bar{r}$ is found by computing the mean of r_{XX} and r_{YY}. Lord (1963) cautioned users of this formula to remember that it requires the assumption that $S^2_X = S^2_Y$.

The difference scores used in Gulliksen's formula were usually computed by subtracting an examinee's score on Test A from his score on Test B when A and B are composed of different test items. Change scores, however, are usually difference scores computed when the same test is administered to an individual on two separate occasions. For this reason Webster (1963) indicated that Formula (6) may be unsatisfactory for computing change score reliability. He noted that this formula derivation rests upon the assumption that errors of measurement are completely uncorrelated, but maintained that this assumption may be unrealistic when the same form of a test is administered twice to an individual. By substituting change scores into the familiar formula for Kuder Richardson 20, he derived an expression for change score reliability which does not require this questionable assumption. Furthermore, it does not

require that the test have equal variances at time 1 and time 2. This formula for change score reliability uses data at the item level and is written:

$$r_{DD} = \frac{k}{k - 1}\left(1 - \frac{\bar{X} + \bar{Y} + \sum^{k}(\bar{x}^2 + \bar{y}^2 - 2\overline{xy}) - 2\bar{f}}{S^2_D}\right). \qquad (7)$$

In this expression, f is the number of items scored 1 on both occasions; $\bar{X}$ is the mean of scores at time 1; $\bar{Y}$, the mean of the scores at time 2; $\bar{x}$ is the mean score of all individuals in the group on item x; $\bar{y}$ is the mean for item y; and k is the number of items.

Bereiter (1963) used the general expression of Cronbach's coefficient alpha and, by substituting change item score for traditional item scores, he defined change score reliability as:

$$r_{DD} = 1 - \frac{\sum S^2_{d_i}}{\sum S^2_{d_i} + \sum_{i \neq j} C_{d_i d_j}} \qquad (8)$$

$S^2_{d_i}$ is the variance of the change item scores and $C_{d_i d_j}$ is their covariance. This expression can be shown to be equivalent to Webster's derivation (Formula 7) except for the absence of the factor $\frac{k}{k - 1}$ in the derivation by Bereiter. This computational formula for change score reliability also uses data at the item level.

In addition, it has the advantage of removing the restriction of dichotomously scored test items.

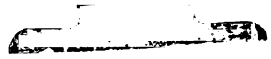
Based upon this formula Bereiter developed a plan for manipulating change score reliability at the item level. He suggested that item analysis techniques could be used to select items which had large change item covariances. When the change item covariances are large for a set of items, there is large variability between subjects on their changes in response to these items. Such variance in change scores results in a lower stability coefficient, r_{XY}, and consequently a higher estimate of r_{DD}. It should be noted that this relationship exists regardless of which formula is used to calculate r_{DD}.

In an empirical study using an attitude questionnaire for college students, Webster and Bereiter (1963) reported that they were able to effect large gains in change score reliability when they employed such an item selection technique. Bereiter, however, makes little mention of the actual index or decision rule used in this item selection.

Horst (1966, p. 387) indicated that most item analysis procedures for raising reliability fall into one of two main categories: correlational and counting procedures. (Counting procedures use response frequency data.) It is obvious that this categorization

scheme can be extended to the realm of change item analysis as well. This classification will be used in the remaining discussion of item analysis procedures designed for selecting items to give reliable change measures.

In 1962, Gruber and Weitman studied changes in item responses to an achievement test. They wanted to measure students' retention of subject matter over time using a pretest-posttest design. Although their goal was not to improve reliability of change scores per se, they suggested that changes in response frequency to items might be useful as a basis for item selection in the measurement of change. In their study the researchers employed two methods of item selection. The first method was based upon observing shifts in response frequency toward a specified, optimal level of difficulty from initial testing to retesting. In the second procedure items were selected on the basis of their pretest response level only. The researchers emphasized the necessity of knowing in advance the direction in which response change is likely to occur over time. The results of this study indicated that it was possible to improve discrimination on the posttest by selecting items on the basis of response shifts. Selecting items on the basis of their pretest difficulty level did not significantly improve the discrimination between subjects on the posttest. However, the authors felt that this could



have been due to the limitations of ceiling effect on their instrument and the small sample size rather than to ineffectiveness of the method itself. No data for the change score reliability was reported, although these estimates could have been easily computed. Thus it is not known if these item selection techniques could have improved the reliability estimate for changes in retention.

The first correlational method of change item analysis was derived by Saupe (1966). Based upon Gulliksen's formula for the correlation between a component element and a composite, Saupe's formula for the correlation for change item with total change score is:

$$r_{dD} = \frac{C_{xX} + C_{yY} - C_{xY} - C_{Xy}}{\sqrt{S^2_x + S^2_y - 2C_{xy}}\ \sqrt{S^2_X + S^2_Y - 2C_{XY}}} \qquad (9)$$

where x and y denote item scores, X and Y are total scores, and C is covariance.

Lord (1968, p. 331) urged that empirical studies be undertaken in this area and stressed the need for development of still other item analysis procedures for change items.

The primary reason that most traditional item analysis methods cannot be used with change scores is due to the nature of the change item itself. As defined by Bereiter (1963), the change

item, $d_i = x_i - y_i$, can have at least three values, 0, 1, or -1. This rules out the possibility of using the biserial correlation which is frequently employed as an index for item selection. (The biserial correlation, point biserial correlation, tetrachoric and phi coefficient all require dichotomously scored items.) Jaspen (1946) developed a formula for the triserial correlation. This was intended to serve as a computational formula for a correlation between two variables when both were assumed to have underlying normal, continuous distributions, but when one distribution had been artificially divided into three categories. This expression is a direct counterpart to the biserial correlation used for computing correlations between a continuous variable and a variable classified into two artificial categories.

Jenkins (1956) presented a simplified version of Jaspen's formula for triserial r:

$$r_{tris} = \frac{M_h y_h + M_m(y_l - y_h) - M_l y_l}{\sigma\left[\frac{y^2_h}{p_h} + \frac{(y_l - y_h)^2}{p_m} + \frac{y^2_l}{p_l}\right]} \qquad (10)$$

where M = mean, y = curve ordinate, and σ = s.d. of scores. The letters h, m, and l represent high, medium, and low categories.

If the item change scores of 1, 0, and -1 are used to designate the divisions of high, medium, and low, it is seen that r_{tris} could be written as an index for change item analysis:

$$r_{tris} = \frac{M_1 y_1 + M_0(y_{-1} - y_1) - M_{-1} y_{-1}}{\sigma \left[\frac{y^2_1}{p_1} + \frac{(y_{-1} - y_1)^2}{p_0} + \frac{y^2_{-1}}{p_{-1}} \right]} \qquad (11)$$

Triserial r, however, had never been used as an index for item selection, despite its availability.

Summary

To summarize this review of literature on item analysis methods for change items and change score reliability, several key points should be noted. First, in recent years there has been great interest in the problems of measuring change and considerable concern over the lack of reliability for change scores. This low reliability made it extremely difficult to predict change for an individual, or to make counseling or placement decisions based on change scores.

Several different formulae for change score reliability were discussed in this chapter. It seems best to conclude that when the same form of a test is used for both initial and final testing, then

Bereiter's Formula (8) or Webster's Formula (7) for change score reliability is preferable to the traditional formula, derived by Gulliksen (Formula 6).

Regardless of which formula is used for r_{DD}, several researchers have suggested that low change score reliability can perhaps be improved through item analysis procedures. Three techniques have been suggested in the literature. These are: selection of items on the basis of observed shifts in response; selection of items on the basis of pretest response frequency when the direction of expected change can be predicted; and use of a correlational index based upon correlation between change item score and total change score.

The development of a formula for triserial correlation was also presented and this statistic was suggested as a fourth possible index for item selection in the measurement of change.

Reports on studies comparing these various change item analysis methods are "conspicuous by their absence" in this review. It is readily apparent that empirical investigation of these methods is essential to determine if they can be successfully used to improve change score reliability. It was toward this end that the study of change item analysis techniques, described in Chapter III, was undertaken.

CHAPTER III

DESIGN OF THE STUDY

An empirical study was designed to compare the change score reliability for subsets of items selected by four different change item analysis procedures. The four procedures compared were Saupe's item-total score correlation for r_{dD}, triserial correlation between change item and change score, selection of items having high variance in change scores, and selection of items on the basis of pretest response frequency. In addition, a control method, selecting items randomly from the original item pool, was used.

The Sample

In the fall of 1958, the first-term freshman class at Michigan State University was tested on a variety of achievement, aptitude, attitude, and personality measures. All freshmen were included in the population who met the following criteria: (1) The student must have been a first time freshman--not a past dropout or a transfer from another university; (2) The student must have been a native born American.

In 1961, a sample was drawn from this original population. This group consisted of students who were still enrolled in the university at that time. These students, then juniors at MSU, were retested on the same measures. The test-retest data from 263 students in this sample were used for this item analysis experiment.

The Instrument

The instrument selected for use in this study was the Inventory of Beliefs, Form I. This attitude survey was developed by the Cooperative Study of Evaluation in General Education under the sponsorship of the American Council on Education Committee on Measurement and Evaluation. The scale was designed to measure an individual's tendency to subscribe to stereotypic beliefs (Lehmann and Dressel, 1963).

Items on this inventory were taken from an original pool of one thousand items, composed by a panel of counselors and evaluation officers from twenty colleges which participated in the Cooperative Study. The 120 statements which were selected for the final scale were written in the form of "pseudo-rational cliches." (American Council on Education, 1953).

Some sample items from this inventory are:

> "No world organization should have the right to tell Americans what they can or cannot do."

"We would be better off if there were fewer psychoanalysts probing and delving into the human mind."

"When things seem black, a person should not complain, for it may be God's will."

"Most Negroes would become overbearing and disagreeable if not kept in their place."

There were four possible responses to each item--Strongly Agree, Agree, Disagree, and Strongly Disagree.

Two separate scoring schemes were used in this study. The scoring instructions from the Instructor's Manual award the examinee with one point for each Disagree or Strongly Disagree response. The second scoring scheme used in the item analysis study awarded one point for a response of Strongly Agree; two points for Agree; three points for Disagree; and four points for Strongly Disagree. Lehmann and Dressel (1963, p. 27) characterized the higher scorer as "mature, flexible, adaptive, and democratic in his relationships with others; a low scorer is immature, rigid in outlook, compulsive, and authoritarian in his relationships with others."

The reliability coefficients reported for this scale in the MSU study ranged from .68 to .95, with a median value of .86 (Lehmann and Dressel, 1963).

The Inventory of Beliefs was considered appropriate for use in this change item analysis study for the following reasons:

1. The scale was designed expressly for the purpose of measuring the attainment of educational objectives. Thus change scores over time were expected to be fairly large and could be meaningfully interpreted.
2. The instrument was professionally developed. Considerable effort went into construction and item analysis. Reliability and validity for this scale had been demonstrated (Dressel and Mayhew, 1952).
3. The internal consistency reliabilities reported were high, but test-retest reliability coefficients after a lapse of time were lower, thus indicating a fairly wide range of individual differences in attitude change.

Design

An item analysis, cross-validation design was employed in this study. The sample was randomly split into two groups. The item analysis group consisted of 132 students; the cross-validation group was composed of 131 students.

Items from the 1958 and 1961 test administrations for these students were scored by both the zero-one scoring method and the one-to-four method described earlier. Item change scores were computed in accordance with Formula (4),

$$d_i = x_i - y_i ,$$

and total change scores were computed for each student.

The data from the item analysis group, scored on the zero-one basis, was submitted to four different item analysis procedures and a control procedure of random selection. Data scored with the one-to-four system was submitted to three item analysis procedures and random selection. (It was necessary to omit the triserial correlation method because it was only appropriate for dichotomously scored items.) Subsets of 15, 30, 60, and 90 items were chosen under each procedure. These item subsets were used for computing reliability estimates for change scores on the cross-validation group data. The actual computation formula for the change score reliability was obtained by substituting Bereiter's definition of change score reliability (Formula 8) into Webster's expression for the Kuder-Richardson 20 for change scores (Formula 7) to get a change score version of Cronbach's coefficient alpha (Cronbach, 1951).

$$r_{DD} = \frac{k}{k - 1} \left[1 - \frac{\sum S^2_{d_i}}{\sum S^2_{d_i} + \sum_{i \neq j} C_{d_i d_j}} \right] \qquad (12)$$

Item Analysis Procedures

Method I was an item analysis procedure based on the variance of the change item scores. After the change item scores, d_i, were computed, the mean change score $\bar{d}_i$ and the change score variance $S^2_{d_i}$ were found for each item. Items with the largest values for $S^2_{d_i}$ were selected. On this basis subsets of 15, 30, 60, and 90 items were chosen from the original set of 120 items.

Method II required that items for the subsets be chosen on the basis of pretest response frequency. With this method it was necessary to take into account the expected direction of the change. Because the Inventory of Beliefs had been developed to measure attainment of objectives of higher education, it seemed reasonable to predict that students' scores would increase over time. (Data from the Lehmann and Dressel study upheld this prediction.) Item means, $\bar{y}_i$, were computed for each item on the pretest (the measure taken in 1958, when the students were freshmen). Items with the lowest mean scores were selected into the 15, 30, 60, and 90 item subsets.

Method III was a correlational item analysis procedure for which the index of item selection was the expression derived by Saupe (Formula 9):

Items

were s

betwee

ing to

Items w

the tes

data th

of 15,

chosen

Testabl

1.

$$r_{dD} = \frac{C_{xX} + C_{yY} - C_{xY} - C_{Xy}}{\sqrt{S^2_x + S^2_y - 2C_{xy}}\ \sqrt{S^2_X + S^2_Y - 2C_{XY}}} .$$

Items which had the greatest correlations with total change score were selected into the test subsets.

For Method IV the triserial correlation coefficients between change item and total change score were computed according to Formula (11):

$$r_{tris} = \frac{M_1 y_1 + M_0(y_{-1} - y_1) - M_{-1} y_{-1}}{\sigma \left[\frac{y^2_1}{p_1} + \frac{(y_{-1} - y_1)^2}{p_0} + \frac{y^2_{-1}}{p_{-1}} \right]} .$$

Items with the highest positive values for r_{tris} were selected into the test subsets. This method, of course, was only applied to the data that had been scored on a zero-one basis on the original tests.

The control method consisted of selecting randomly subsets of 15, 30, 60, and 90 items for comparison with those which had been chosen by the systematic item analysis procedures.

Testable Hypotheses

The specific hypotheses tested in this study were:

1. The mean change score reliability for items chosen by Method III (Saupe's correlation between change item and

2

3.

4.

Statistic

reliabilit

Using the

total change score) would be greater than the mean reliability for item subsets chosen by any other item analysis method or by the control method of random selection.

2. Mean change score reliability for subsets of items chosen by Method IV (triserial correlation) would be greater than the mean reliability for the subsets of items chosen by either the response frequency methods or by random selection.
3. Mean change score reliability for the subsets of items selected by Method I (using change item variance) would be greater than the mean reliability of item subsets chosen by pretest response frequency or by random selection.
4. Mean change score reliability for the subsets of items selected by Method II (using pretest response frequency) would be greater than the mean reliability of subsets of randomly selected items.

Statistical Analysis

Two procedures were used to compare the change score reliability coefficients computed on the cross-validation sample. Using the first method, all reliability estimates for each subset of

item

stude

son.

ences

valida

one sa

then ra

score

of the

with a

of item

Tukey'

was us

of the

Summa

gan St

an att

been c

and Dr

items were computed on the whole cross-validation sample of 131 students. This was to provide the best overall descriptive comparison.

To test the statistical significance of the observed differences in the change score reliability coefficients, the cross-validation group was divided into smaller independent samples--one sample for each item analysis method. These samples were then randomly assigned to the item analysis procedures and change score reliabilities were computed. Fisher r-to-Z transformations of the reliability coefficients were used, and the values were analyzed with a two-way analysis of variance. (One main effect was method of item analysis; the other was number of items in the subset.) Tukey's test for an honestly significant difference (Kirk, 1968, p. 88) was used to test the significance of the differences between the means of the reliability estimates.

Summary

Test-retest data were obtained from a sample of 263 Michigan State University students in their freshman and junior years on an attitude survey called the Inventory of Beliefs. (These data had been collected as part of a longitudinal study conducted by Lehmann and Dressel from 1958 to 1962.)

The data were scored by two different methods--a zero-one scoring method and a one-to-four scaling method. Item change scores were computed for all 120 items on the questionnaire.

Data from half of the sample were subjected to four different item analysis procedures and a control procedure of random selection. The item analysis procedures used for items scored zero-one were: Saupe's correlation index, triserial correlation, selection for large change variance, and selection on the basis of pretest response frequency. All of these same procedures were used for the data scored on a one-to-four scale, except for triserial correlation. Subsets of 15, 30, 60, and 90 items were chosen by each method.

Change score reliabilities for these subsets of items were computed using change score data from the cross-validation group. A change score reliability version of coefficient alpha was used. A two-way analysis of variance and a Tukey post hoc comparison test were used to test for differences in change-score reliability for the items chosen by different methods.

Result

a one-

control

of 15,

analys

on iter

change

item a

from

relial

respo

relia

Saupe

relial

CHAPTER IV

RESULTS

Results for One-to-Four Scoring

When the items of the Inventory of Beliefs were scored on a one-to-four scale, three methods of change item analysis and a control method of random selection were employed to select subsets of 15, 30, 60, and 90 items. The three methods of change item analysis were: selection on pretest response frequency, selection on item change score variance, and Saupe's correlation between change item score and total change score. Detailed results of the item analyses are presented in the Appendix.

After the subsets of items had been selected, using data from the 132 students in the item analysis group, the change score reliability for each item subset was computed using the item responses of the 131 students in the cross-validation group. These reliability coefficients are presented in Table 4.1.

From the results presented in Table 4.1, it is apparent that Saupe's method of change item analysis consistently resulted in more reliable subsets of items than did either of the other two item analysis

methods or the control method of random selection. There was little difference between the reliability coefficients of item subsets chosen by the two response frequency methods (Method I and Method II); however, both of these methods resulted in higher reliability of change scores than did the control method for subsets of 15, 30, 60, and 90 items.

TABLE 4.1. --Change score reliability coefficients computed for the total cross-validation sample using the one-to-four scoring system.

Item Analysis Method	Number of Items			
	15	30	60	90
Method I (Change Variance)	.50	.61	.75	.83
Method II (Pretest Frequency)	.50	.65	.78	.83
Method III (Saupe's r_{dD})	.63	.70	.80	.85
Method IV (Random)	.30	.49	.70	.80

Another point that should be noted from the data presented in Table 4.1 is that the differences between reliability coefficients were greater when fewer items were selected from the original pool. At the 90-item level the reliability values ranged only from .85 for Method III (Saupe's) to .80 for the control. At the 15-item level, however, the range was from .63 for Saupe's method to .30 for the control.

To test the statistical significance of the differences between change score reliability estimates obtained for item subsets chosen by the different methods, the cross-validation sample was divided into four random subsamples with 32 students in each group. Each of these samples was then randomly assigned to a different item analysis method. The reliability coefficients for 15, 30, 60, and 90 items chosen by a method were then computed using the data from the small group which had been assigned to it. Thus reliability estimates obtained under different item analysis methods were calculated for independent samples to meet the assumptions of the analysis of variance model. These change score reliability coefficients are reported in Table 4.2.

TABLE 4.2. --Change score reliability coefficients computed for independent cross-validation samples using the one-to-four scoring system.

Item Analysis Method	Number of Items			
	15	30	60	90
Method I (Change Variance) Sample 1	.48	.61	.76	.84
Method II (Pretest Frequency) Sample 2	.56	.67	.76	.80
Method III (Saupe's r_{dD}) Sample 3	.76	.80	.86	.89
Method IV (Random) Sample 4	.36	.42	.64	.76

Fisher r-to-Z transformations of the values in Table 4.2 were used as the dependent variables in a two-way analysis of variance (fixed effects model) with one observation per cell (Winer, 1962, p. 217). In this analysis, item analysis method was one independent factor with four levels; number of items was the second factor with four repeated measures on each sample. Because there was only one replication per cell, a Tukey one-degree-of-freedom test for nonadditivity (Winer, 1962, p. 218) was conducted to test for the confounding effects of an interaction in the error term prior to running the two-way ANOVA. No significant interaction effect was detected at the alpha level of .05.

TABLE 4.3. --Two-way analysis of variance for the effects of item analysis method and number of items on change score reliability (with the one-to-four scoring system).

Source of Variance	Sums of Squares	d.f.	M.S.	F Ratio
Item Analysis Method	.622	3	.207	51.75**
Number of Items	.742	3	.247	61.75**
Residual	.039	9	.004	
Total	1.403	15		

**Significant at alpha = .01.

Fisher r-to-z transformations of the values in Table 4.2 were used as the dependent variables in a two-way analysis of variance (fixed effects model) with one observation per cell (Winer, 1962, p. 217). In this analysis, item analysis method was one independent factor with four levels; number of items was the second factor with four repeated measures on each sample. Because there was only one replication per cell, a Tukey one-degree-of-freedom test for nonadditivity (Winer, 1962, p. 218) was conducted to test for the confounding effects of an interaction in the error term prior to running the two-way ANOVA. No significant interaction effect was detected at the alpha level of .05.

TABLE 4.3.--Two-way analysis of variance for the effects of item analysis method and number of items on change score reliability (w/, the one-day scoring system).

Source of Variance	SS	df	MS	F Ratio
Item Analysis Method	.024	3	.207	31.75**
Number of Items	.142	3	.247	61.75**
Residual	.028	9	.004	
Total	1.401	15		

**Significant at alpha = .01

Results of the analysis of variance are presented in Table 4.3. The main effect for number of items was significant at the alpha level of .01, using a conservative F test with 1 and 3 degrees of freedom. Main effect for item analysis method was significant at the alpha level of .01, using an F test with 3 and 3 degrees of freedom (Greenhouse and Geisser, 1959).

Testing Hypotheses for One-to-Four Scoring

The hypotheses tested were:

1. The mean change score reliability for item subsets chosen by Method III (Saupe's r_{dD}) would be greater than the mean reliabilities for subsets of items chosen by any other item analysis method or by the control method of random selection.
2. Mean change score reliability for subsets of items chosen by Method I, using change item variance, would be greater than the mean reliability of item subsets chosen by pretest response frequency or by random selection.
3. Mean change score reliability for the subsets of items selected by Method II, using pretest response frequency, would be greater than the mean reliability of subsets of randomly selected items.

Results of the analysis of variance are presented in Table 4.3. The main effect for number of items was significant at the alpha level of .01, using a conservative F test with 1 and 3 degrees of freedom. Main effect for item analysis method was significant at the alpha level of .01, using an F test with 3 and 3 degrees of freedom (Greenhouse and Geisser, 1959)

Testing Hypotheses for
One-to-Four Scoring

The hypotheses tested were:

1. The mean change score reliability for item subsets chosen by Method III (Sample r_{III}) would be greater than the mean reliabilities for subsets of items chosen by any other item analysis method using the control method of random selection.

2. Mean change score reliability for subsets of items chosen by Method I, using item variance, would be greater than the mean reliability of item subsets chosen by pretest response frequency or by random selection.

3. Mean change score reliability for the subsets of items selected by Method II, using pretest response frequency, would be greater than the mean reliability of subsets of randomly selected items.

To test these hypotheses, post hoc comparisons were made to determine the significance of the differences between the mean reliability values obtained under the different item analysis methods. A multiple comparison test for making a series of pairwise comparisons, developed by Tukey, was employed (Kirk, 1968, p. 88). An HSD (honestly significant difference) value was computed in accordance with the formula:

$$HSD = q_{\gamma}\sqrt{\frac{MS_{error}}{n}} \qquad (13)$$

where n is the number of levels or treatments, q is a value obtained from the tabled distribution of the studentized range statistic, and γ is the number of degrees of freedom associated with the error term.

Differences between the mean reliabilities for item subsets selected by the various methods of item analysis are presented in Table 4.4. (Reliability estimates were converted to Fisher r-to-Z transformations for testing.)

As Table 4.4 shows, the first hypothesis is supported. Change score reliability for subsets of items chosen by Saupe's method is significantly greater than that of subsets of items selected by any other method.

The first part of the second hypothesis is not supported. The reliability for sets of change items selected for their variance

To test these hypotheses, post hoc comparisons were made to determine the significance of the differences between the mean reliability values obtained under the different item analysis methods. A multiple comparison test for making a series of pairwise comparisons, developed by Tukey, was employed (Kirk, 1968, p. 88). An HSD (honestly significant difference) value was computed in accordance with the formula:

$$HSD = q\sqrt{\frac{MS_{error}}{n}} \qquad (13)$$

where n is the number of levels or treatments, q is a value obtained from the tabled distribution of the studentized range statistic, and $\sqrt{}$ is the number of degrees of freedom associated with the error term.

Difference between two mean reliabilities for item subsets selected by the various methods of item analysis are presented in Table 4.4. (Reliability estimates were converted to Fisher r-to-Z transformations for testing.)

As Table 4.4 shows, the first hypothesis is supported. Change score reliability for subsets of items chosen by Super's method is significantly greater than that of subsets of items selected by any other method.

The first part of the second hypothesis is not supported. The reliability for sets of change items selected for their variance

is not better than reliability for items chosen on the basis of pretest response frequency; it is, however, significantly greater than the reliability of the randomly selected subsets of items.

TABLE 4.4. --Differences between reliability estimates for items chosen by different item analysis methods (one-to-four scoring).

	Change Variance	Pretest Frequency	Saupe r_{dD}
Change Variance		.018	.328*
Pretest Frequency			.310*
Saupe r_{dD}			
Random	.226*	.244*	.554**

*Significant at alpha = .05, HSD = .218.

**Significant at alpha = .01, HSD = .389.

The third hypothesis is also supported by the data. Items can be chosen on the basis of pretest response frequency which have higher change score reliability than an equal number of items randomly chosen.

Results for Zero-One Scoring

When the items on the attitude survey were scored on a zero-one basis, it was possible to introduce a fifth method of item

is not better than reliability for items chosen on the basis of pretest response frequency; it is, however, significantly greater than the reliability of the randomly selected subsets of items.

TABLE 4.4.--Differences between reliability estimates for items chosen by different item analysis methods (one-to-four scoring).

	Change Variance	Pretest Frequency	Snape r_{AD}
Change Variance		.018	.328*
Pretest Frequency			.310*
Snape r_{AD}			
Random	.328*	.344*	.254**

*Significant at about .01 HSD = .218.

**Significant at about .01 HSD = .329.

The third hypothesis is also supported by the data. Items can be chosen on the basis of pretest response frequency which have higher change score reliability than an equal number of items randomly chosen.

Results for Zero-One Scoring

When the items on the attitude survey were scored on a zero-one basis, it was possible to introduce a fifth method of item

selection (triserial correlation) in addition to the three-item analysis methods used for one-to-four scoring and random selection. The change score reliabilities for the 15, 30, 60, and 90 item subsets were computed using the responses of the entire cross-validation sample. These change score reliability estimates are presented in Table 4.5. The differences between the methods of item analysis were much less pronounced under this scoring system. In general, however, all four methods of change item analysis consistently resulted in higher estimates of change score reliability than did the technique of random selection. The greatest differences, again, were observed when fewer items were selected from the original pool.

TABLE 4.5. --Change score reliability coefficients computed for the total cross-validation sample using the zero-one scoring method.

Item Analysis Method	Number of Items			
	15	30	60	90
Method I (Change Variance)	.52	.56	.68	.72
Method II (Pretest Frequency)	.36	.52	.67	.72
Method III (Saupe's r_{dD})	.33	.49	.68	.74
Method IV (Triserial r)	.37	.56	.68	.75
Method V (Random)	.21	.48	.57	.67

selection (Pearsonian correlation) in addition to the three item analysis methods used for one-to-four scoring and random selection. The change score reliabilities for the 15, 30, 60, and 90 item subsets were computed using the responses of the entire cross-validation sample. These change score reliability estimates are presented in Table 4.5. The differences between the methods of item analysis were much less pronounced under this scoring system. In general, however, all four methods of change item analysis consistently resulted in higher estimates of change score reliability than did the technique of random selection. The greatest difference, again, were observed when fewer items were selected from the original pool.

TABLE 4.5.--Change score reliability coefficients computed for the total cross-validation sample using the zero-one scoring method

Item Analysis Method	Number of Items			
	15	30	60	90
Method I (Change Variance)	.52	.65	.78	.72
Method II (Pearsonian)	.56	.73	.67	.72
Method III (Sample r_{pb})	.53	.69	.78	.74
Method IV (Tetrachoric)	.57	.68	.78	.75
Method V (Random)	.31	.48	.57	.67

To test the statistical significance of the differences between the change score reliability estimates of items selected by the various item analysis methods, the cross-validation group was divided into five independent samples with 26 students per sample. Each sample was then randomly assigned to be used for calculating the reliabilities for 15, 30, 60, and 90 items chosen by a particular item analysis method. The change score reliability coefficients obtained on these independent samples are reported in Table 4.6.

TABLE 4.6. --Change score reliability coefficients computed for independent cross-validation samples using the zero-one scoring system.

Item Analysis Method	Number of Items			
	15	30	60	90
Method I (Change Variance) Sample 1	.50	.67	.72	.76
Method II (Pretest Frequency) Sample 2	.48	.60	.68	.75
Method III (Saupe's r_{dD}) Sample 3	.32	.44	.65	.74
Method IV (Triserial r) Sample 4	.45	.62	.71	.76
Method V (Random) Sample 5)	.11	.35	.60	.74

To test the statistical significance of the differences between the change score reliability estimates of items selected by the various item analysis methods, the cross-validation group was divided into five independent samples with 26 students per sample. Each sample was then randomly assigned to be used for calculating the reliabilities for 10, 30, 50, and 90 items chosen by a particular item analysis method. The change score reliability coefficients obtained on these independent samples are reported in Table 4.6.

TABLE 4.6.--Change score reliability coefficients computed for independent cross-validation samples using the zero-one scoring system

Item Analysis Method	Number of Items			
	10	30	50	90
Method I (Change Variance) Sample 1	.76	.77	.77	.76
Method II (Pretest Proficiency) Sample 2	.44	.70	.68	.79
Method III (Sauser's $r_{(2)}$) Sample 3	.72	.84	.85	.74
Method IV (Triserial η) Sample 4	.75	.82	.71	.76
Method V (Random) Sample 5	.11	.35	.60	.74

A two-way analysis of variance was performed using Fisher r-to-Z transformations of the reliability coefficients in Table 4.6. Prior to running the ANOVA, a Tukey one-degree-of-freedom test for nonadditivity was conducted to detect the significance of an interaction effect. No significant interaction effect was found at the alpha level of .05. Results of the two-way ANOVA are presented in Table 4.7.

TABLE 4.7. --Two-way analysis of variance for the effects of item analysis method and number of items on change score reliability (with zero-one scoring).

Source of Variance	Sums of Squares	d.f.	M.S.	F Ratio
Item Analysis Method	.219	4	.055	7.857*
Number of Items	.914	3	.305	43.570**
Residual	.085	12	.007	
Total	1.218	19		

*Significant at alpha = .05.

**Significant at alpha = .01.

Using the conservative F-test with 4 and 4 degrees of freedom, the main effect of item analysis method was significant at the alpha level .05. The effect of number of items was significant at the alpha level of .01, using a conservative F-test with 1 and 4 degrees of freedom.

A two-way analysis of variance was performed using Fisher r-to-z transformations of the reliability coefficients in Table 4.6. Prior to running the ANOVA, a Tukey one-degree-of-freedom test for nonadditivity was conducted to detect the significance of an interaction effect. No significant interaction effect was found at the alpha level of .05. Results of the two-way ANOVA are presented in Table 4.7.

TABLE 4.7.--Two-way analysis of variance for the effects of item analysis method and number of items on change score reliability (with zero-one scoring)

Source of Variance	Sums of Squares	d.f.	M.S.	F Ratio
Item Analysis Method	.012	1	.005	7.857*
Number of Items	.054	1	.030	43.370**
Residual	.007	12	.001	
Total	.121	14		

*Significant at alpha = .05.

**Significant at alpha = .01.

Using the conservative F-test with 1 and 4 degrees of freedom, the main effect of item analysis method was significant at the alpha level of .05. The effect of number of items was significant at the alpha level of .01, using a conservative F-test with 1 and 4 degrees of freedom.

Testing Hypotheses for Zero-One Scoring

When the items were scored on a zero-one system there were four hypotheses of interest.

1. The mean change score reliability for items chosen by Method III (Saupe's correlation) would be greater than the mean reliability of item subsets chosen by any other item analysis method or by random selection.
2. Mean change score reliability for subsets of items chosen by Method IV (triserial correlation) would be greater than the mean reliability for the subsets of items chosen by the response frequency methods or by random selection.
3. Mean change score reliability for the subsets of items selected by Method I (using change variance) would be greater than the mean reliability of item subsets chosen by frequency of pretest responses or by random selection.
4. Mean change score reliability for the subsets of items selected by Method II (using pretest response frequency) would be greater than the mean reliability of subsets of randomly selected items.

A post hoc comparison test for differences between means was employed to test these hypotheses. Tukey's test for an honestly

Testing Hypotheses for
Zero-One Scoring

When the items were scored on a zero-one system there were four hypotheses of interest.

1. The mean change score reliability for items chosen by Method III (Saupe's correlation) would be greater than the mean reliability of item subsets chosen by any other item analysis method or by random selection.

2. Mean change score reliability for subsets of items chosen by Method IV (triserial correlation) would be greater than the mean reliability for the subsets of items chosen by the response frequency methods or by random selection.

3. Mean change score reliability for the subsets of items selected by Method I (mean change on items) would be greater than the mean reliability of item subsets chosen by frequency of pivotal responses or by random selection

4. Mean change score reliability for the subsets of items selected by Method II (using pivotal response frequency) would be greater than the mean reliability of subsets of randomly selected items.

A post hoc comparison test for differences between means was employed to test these hypotheses. Tukey's test for an honestly

significant difference was used for making the pairwise comparisons between means. Results of these comparisons are reported in Table 4.8. Three methods of item analysis were significantly better than random selection in producing reliable change scales. These were: selecting on change item variance, selecting on pretest response frequency, and triserial correlation. Saupe's correlational method did not produce results that were significantly better than random selection. Only one significant difference was found between the item analysis methods themselves. Selection of items on the basis of change variance was found to yield higher mean change reliability than selection on the basis of Saupe's r_{dD}.

TABLE 4.8. --Differences between mean change score reliabilities for items chosen by different methods. (Scores are Fisher r-to-Z transforms.)

	Change Variance	Pretest Frequency	Saupe r	Triserial r
Change Variance				
Pretest Frequency	.057			
Saupe's r_{dD}	.178	.121		
Triserial r	.041	-.016	-.137	
Random	.285*	.228*	.125	.244*

*Significant at alpha = .05, HSD = .201.

significant difference was used for making the pairwise comparisons between means. Results of these comparisons are reported in Table 4.8. Three methods of item analysis were significantly better than random selection in producing reliable change scales. These were: selecting on change item variance, selecting on pretest response frequency, and triserial correlation. Soupir's correlational method did not produce results that were significantly better than random selection. Only one significant difference was found between the item analysis methods themselves. Selection of items on the basis of change variance was found to yield higher mean change reliability than selection on the basis of Soupir's r_{bis}.

TABLE 4.8. --Differences between mean change score reliabilities for items chosen by different methods. (Scores are Fisher's z-transformed.)

	Change Variance	Pretest Frequency	Soupir's r_{bis}	Triserial
Change Variance				
Pretest Frequency	.17			
Soupir's r_{bis}	.139	.321		
Triserial	.041	- .098	-.137	
Random	.284*	.238*	.120	.244*

*Significant at alpha = .05, LSD = .201.

Thus the first hypothesis is not supported. Saupe's method of change item analysis is not superior to other item analysis methods, nor is it better than random selection, in choosing items for reliable change scales.

The second hypothesis is partially supported. Triserial correlation is better than random selection, but is not superior to response frequency methods for selecting change items.

The third hypothesis is also only supported by the fact that using change item variance as an index for item selection is better than random selection. This method, however, is not significantly better than selecting items on the basis of pretest response frequency.

The fourth hypothesis is upheld. Items can be chosen on the basis of pretest response which have significantly higher change score reliabilities than items which are randomly chosen.

Summary

Results of the two-way analysis of variance and Tukey post hoc comparisons showed that for the one-to-four item scoring system:

1. Saupe's r_{dD} was superior to random selection and to both selection for change score variance and selection on pre-test response frequency.

Thus the first hypothesis is not supported. Saupe's method of change item analysis is not superior to other item analysis methods, nor is it better than random selection, in choosing items for reliable change scales.

The second hypothesis is partially supported. Triserial correlation is better than random selection, but is not superior to response frequency methods for selecting change items.

The third hypothesis is also only supported by the fact that using change item variance as an index for item selection is better than random selection. This method, however, is not significantly better than selecting items on the basis of pretest response frequency.

The fourth hypothesis is supported. Items can be chosen on the basis of pretest response which are significantly higher change score reliabilities than items randomly chosen

Summary

Results of the two-way analysis of variance and Tukey post hoc comparisons showed that for the one-to-four item scoring system:

1. Saupe's r_{dD} was superior to random selection and to both selection for change score variance and selection on pre-test response frequency.

2. Both pretest response frequency and the change variance methods were better than random selection for choosing reliable change item subscales.
3. Selection on change item variance was no better than selection on pretest response frequency in providing reliable change scales.

When the items were scored dichotomously, the results showed that:

1. The methods of triserial correlation, selection for change variance, and selection on pretest response frequency were all superior to random selection of items for change scales.
2. There were no significant differences between these successful methods of item analysis.
3. Saupe's r_{dD} was not significantly better than random selection of items for measuring change.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

In recent years researchers have become increasingly interested in the problems of measuring change. Low change score reliability has presented a particularly challenging problem to researchers in this area. Bereiter (1963) suggested that item analysis techniques could be applied to change items in an attempt to improve change score reliability.

A review of the literature revealed that several techniques for change item analysis were available; however, there was a dearth of empirical research to demonstrate the effectiveness of these procedures or to compare their ability to increase change score reliability. The four methods of item analysis suitable for change items were: selection on the basis of change item score variance; selection on the basis of pretest response frequency; selection on Saupe's correlation between change item score and total score; and selection on triserial correlation. (The latter method

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

In recent years researchers have become increasingly interested in the problem of measuring change. Low change score reliability has presented a particularly challenging problem to researchers in this area. Bryan (1963) suggested that item analysis techniques could be applied to change items in an attempt to improve change score reliability.

A review of the literature revealed that several techniques for change item analysis were available. However, there was a dearth of empirical research to substantiate the effectiveness of these procedures or to compare their ability to increase change score reliability. The four purposes of item analysis suitable for change items were: selection on the basis of change item score variance; selection on the basis of pretest response frequency; selection on Soper's correlation between change item score and total score; and selection on Tucker correlation. (The latter method

was restricted to the case where items were dichotomously scored on each occasion.)

An empirical study was undertaken to determine whether these methods of change item analysis could lead to the selection of more reliable subsets of items than could items chosen by random selection. Comparisons between the various methods were also made. The sample used for item analysis and cross-validation was a group of 263 students at Michigan State University who had been tested on the Inventory of Beliefs as freshmen in 1958, and who were retested on this attitude survey in 1961.

Half of this sample were assigned to an initial item analysis group. On the basis of their responses the four item analysis procedures were carried out and subsets of 15, 30, 60, and 90 items were selected by each procedure from the original pool of 120 items. In addition, a control procedure of random selection was also used to choose item subsets.

The items selected by the item analysis procedures were then scored for the cross-validation group. Two change score reliabilities were calculated from these responses. First, all reliability estimates were computed for the entire group of 131 students. Secondly, the cross-validation group was divided into smaller independent samples and reliabilities for item subsets chosen

by different methods were computed on independent samples. Hypotheses were tested by using a two-way analysis of variance with post hoc comparisons.

The results of the analysis showed that when the items were scored on a one-to-four scale, the three methods of item analysis used resulted in significantly higher change score reliability than did random selection. Saupe's r_{dD} was the most successful in producing high change score reliability. Selection on the basis of pretest frequency and change score variance were equally effective.

When the items were scored on a zero-one basis, three methods of item analysis resulted in greater change score reliability than did random selection--selection on change variance, selection on pretest response frequency, and triserial correlation. One method (Saupe's r_{dD}) proved to be no better than random selection. No significant differences were found between the three methods which were successful in improving change score reliability; they were equally effective.

Conclusions

From this comparative study of change item analysis techniques, several conclusions can be drawn.

1. It is possible to produce more reliable instruments for measuring individual change if items are selected through

by different methods were computed on independent samples.

Hypotheses were tested by using a two-way analysis of variance with post hoc comparisons.

The results of the analysis showed that when the items were scored on a one-to-four scale, the three methods of item analysis used resulted in significantly higher change score reliability than did random selection. Saupe's r_{Qb} was the most successful in producing high change score reliability. Selection on the basis of pretest frequency and change score variance were equally effective.

When the items were scored on a zero-one basis, three methods of item analysis resulted in greater change score reliability than did random selection. Selection on Saupe's r_{Qb}, selection on pretest response frequency, and biserial correlation. One method (Saupe's r_{Qb}) proved to be better than random selection. No significant differences were found between the three methods which were successful in increasing change score reliability; they were equally effective.

Conclusions

From this comparative study of change item analysis techniques, several conclusions can be drawn.

1. It is possible to produce more reliable instruments for measuring individual change if items are selected through

the systematic item analysis procedures suggested in this study.

2. When a wide range of item responses is permitted (such as when items are scored on a one-to-four or one-to-five scale), the methods recommended for use are Saupe's r_{dD}, selection on the basis of large change variance, and selection on the basis of pretest response frequency.
3. When the range of item responses is restricted to a dichotomy, the recommended methods are selection on the basis of large change variance, selection on pretest response frequency, and triserial correlation.

Discussion

It should be noted that the differences between the item analysis methods and between item analysis methods and random selection were more dramatic when a smaller number of items was chosen. This probably would indicate that change item analysis techniques would be most useful when a small portion of items are chosen from a larger original pool. It appears, however, to contradict that fact that no significant interaction was found between number of items selected and the method of selection. In view of this it is likely that a Type II error occurred in the Tukey

the systematic item analysis procedures suggested in this study.

2. When a wide range of item responses is permitted (such as when items are scored on a one-to-four or one-to-five scale), the methods recommended for use are Guilford's "D" selection on the basis of large change variance, and selection on the basis of greatest response frequency.

3. When the range of item responses is restricted to a dichotomy, the recommended methods are selection on the basis of large change variance, selection on greatest response frequency, and triserial correlation.

Discussion

It should be noted [illegible] between the item analysis methods and [illegible] analysis methods and random selection were more discriminating [illegible] number of items was chosen. This probably [illegible] item analysis techniques would be most useful using a small portion of items are chosen from a large original pool. It appears, however, to contradict that fact that no significant interaction was found between number of items selected and the method of selection. In view of this it is likely that a type II error occurred in the Tukey

one-degree-of-freedom test for interaction. Even if this were the case, the results of the analysis-of-variance can be accepted with confidence since the presence of an undetected interaction would have resulted in an overestimate of error variance and, hence, a more conservative statistical test for the main effects of number of items and method of selection.

Another point that should not be overlooked is that, statistically speaking, most of the items on this scale functioned effectively. Less than ten items were found which had negative triserial r or r_{dD} values. It is not unreasonable to speculate that if there had been a higher percentage of poor items on the test, the item analysis techniques might have worked even better, and/or differences between the techniques might have been more apparent.

Theoretically, it is not surprising that differences between change item analysis techniques seemed to be greater when the one-to-four scoring scheme was used. When there was a greater possible range of response, there was a greater possible range for the variances of the change scores and, consequently, a greater possible range for change item covariances. Thus a better distinction could be made between "good" and "bad" items. Under the dichotomous scoring system, the items tended to appear more similar with regard to their change variances and intercorrelations.

one-degree-of-freedom test for interaction. Even if this were the case, the results of the analysis-of-variance can be accepted with confidence since the presence of an undetected interaction would have resulted in an overestimate of error variance and, hence, a more conservative statistical test for the main effects of number of items and method of selection.

Another point that should not be overlooked is that, statistically speaking, most of the items on this scale functioned effectively. Less than ten items were found which had negative triserial r_{Db} values. It is not unreasonable to speculate that if there had been a higher percentage of poor items on the test, the item analysis techniques might have worked even better, and/or differences between the techniques might have been more apparent.

Theoretically, it is not surprising that differences between change item analysis techniques seemed to be greater when the one-to-four scoring scheme was used. When there was a greater possible range of response, there was a greater possible range for the variances of the change scores and, consequently, a greater possible range for change item covariances. Thus a better distinction could be made between "good" and "bad" items. Under the dichotomous scoring system, the items tended to appear more similar with regard to their change variances and intercorrelations.

From a practitioner's viewpoint, several of the findings of this study can be applied to the area of constructing instruments to measure change. First, it appears that change item analysis can be a profitable approach to solving the change score reliability problem. Certainly researchers should consider using these techniques when constructing new instruments to measure change or when forced to shorten an already existing scale.

Second, the use of a multiple-response format for items seems to allow a more sensitive observation of change and makes the selection of a method of change item analysis an important consideration.

A third point, having great significance for the longitudinal researcher, is that considerable time and expense might be saved by selecting items on the basis of pretest response alone. It should be remembered, however, that this can only be done when the direction of change can be predicted in advance.

Implications for Future Research

This study has been somewhat of a "pioneer exploration" into the area of change item analysis. It has revealed, nonetheless, that empirical research on techniques for selecting items to measure

change can be a profitable venture yielding useful information for test construction.

Only four possible methods of change item analysis were compared in this study. As other methods are developed, they should be systematically compared with these techniques. Two possible new methods which could be considered are factor analysis and multiserial correlation. Change item scores could be subjected to factor analysis and chosen on the basis of their factor loadings, just as regular items are often selected. This requires, however, a much larger sample size than was employed in this study. Lange (1969) found that factor loadings for 40 items were unstable with a sample size less than 400. Another method which could prove useful would be the use of the general multiserial correlation developed by Jaspen (1946). The triserial r employed in this study was a simplified version of Jaspen's multiserial correlation formula. The general formula could be expanded to render correlation coefficients for total change score and change item responses scored on a one-to-four or a one-to-five scale.

Also, before the findings of this study can be generally accepted, replication is needed using other populations and other instruments.

change can be a profitable venture yielding useful information for test construction.

Only four possible methods of change-item analysis were compared in this study. As other methods are developed, they should be systematically compared with these techniques. Two possible new methods which could be considered are factor analysis and multiserial correlation. Change item scores could be subjected to factor analysis and chosen on the basis of their factor loadings, just as regular items are often selected. This requires, however, a much larger sample size than was employed in this study. Lange (1969) found that factor loadings for 10 items were unstable with a sample size less than 400. Another method which could prove useful would be the use of the general multiserial correlation developed by Jaspen (1965). The [illegible] employed in this study was a simplified version of Jaspen's [illegible] correlation formula. The general formula could be [illegible] to render correlation coefficients for total change scores and change item responses scored on a one-to-four or a one-to-five [illegible].

Also, before the findings of this study can be generally accepted, replication is needed using other populations and other instruments.

An actual study of the content of the change items themselves was not within the scope of this investigation. It remains a very important, but, as yet, unexplored area. Cox (1965) warned test constructors to remember that item selection on statistical criteria alone might change the nature of the test by eliminating items designed to measure a specific objective. He proposed a method to use in conjunction with statistical item analysis to insure that items were maintained in the test to cover all vital objectives of the evaluation. Such selection of items to measure objectives could be practiced equally well with change items now that feasible, statistical item selection techniques are available.

A final implication of this study goes beyond the area of constructing instruments to the broader area of measuring change. This study has demonstrated that researchers need no longer fear to undertake a study of individual change because of the insurmountable problem of low change score reliability. Through item analysis, instruments can be developed which will provide reliable measures of individual change.

An actual study of the content of the change items themselves was not within the scope of this investigation. It remains a very important, but, as yet, unexplored area. Cox (1965) warned test constructors to remember that item selection on statistical criteria alone might change the nature of the test by eliminating items designed to measure a specific objective. He proposed a method to use in conjunction with statistical item analysis to insure that items were maintained in the test to cover all vital objectives of the evaluation. Such selection of items to measure objectives could be practiced equally well with change items now that feasible, statistical item selection techniques are available.

A final implication of this study goes beyond the area of constructing instruments to the broader area of measuring change. This study has demonstrated that researchers need no longer fear to undertake a study of individual change because of the insurmountable problem of low change score reliability. Through item analysis, instruments can be developed which will provide reliable measures of individual change.

BIBLIOGRAPHY

BIBLIOGRAPHY

American council on Education Committee on Measurement and Evaluation. Instructor's manual for the inventory of beliefs. Washington, D.C., 1953.

Bereiter, Carl M. Some persisting dilemmas in the measurement of change. Chapter 1 in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Cox, Richard C. Item selection techniques and evaluation of instructional objectives. Journal of Educational Measurement, 1965, 2, 181-185.

Cronbach, Lee J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Dressel, Paul L., and Mayhew, Lewis B. General education explorations in evaluation. Washington, D.C.: American Council on Education, 1954.

Greenhouse, S. W., and Geisser, S. On methods in the analysis of profile data. Psychometrika, 1959, 24, 92-112.

Gruber, H. E., and Weitman, M. Item analysis and the measurement of change. Journal of Educational Research, 1962, 6, 287-289.

Gulliksen, Harold. Theory of mental tests. New York: Wiley, 1950.

Horst, Paul. Multivariate models for evaluating change. Chapter 6 in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

BIBLIOGRAPHY

American Council on Education Committee on Measurement and Evaluation. Instructor's manual for the inventory of beliefs. Washington, D. C., 1953.

Bereiter, Carl M. Some persisting dilemmas in the measurement of change. Chapter 1 in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Cox, Richard C. Item selection techniques and evaluation of instructional objectives. Journal of Educational Measurement, 1965, 2, 181-185.

Cronbach, Lee J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Dressel, Paul L., and Mayhew, Lewis B. General education explorations in evaluation. Washington, D. C.: American Council on Education, 1954.

Greenhouse, S. W., and [illegible] of the methods in the analysis of profile data. [illegible], 1959, 24, 95-112.

Gruber, H. E., and Weitman, M. Error analysis and the measurement of change. Journal of Experimental Research, 1962, 6, 287-293.

Gulliksen, Harold. Theory of mental tests. New York: Wiley, 1950.

Horst, Paul. Multivariate models for evaluating change. Chapter 8 in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Horst, Paul. Psychological measurement and prediction. Belmont, California: Wadsworth, 1966.

Jaspen, Nathan. Serial correlation. Psychometrika, 1946, 11, 23-30.

Jenkins, William L. Triserial r--a neglected statistic. Journal of Applied Psychology, 1956, 40, 63-64.

Kirk, Roger E. Experimental design procedures for the behavioral sciences. Belmont, California: Wadsworth, 1968.

Lange, Allan L. An empirical study of sampling error in factor analysis. Unpublished doctoral dissertation, Michigan State University, 1969.

Lehmann, Irvin J., and Dressel, Paul L. Changes in critical thinking, attitudes, and values associated with college attendance. Final Report of Cooperative Research Project No. 1646. East Lansing: Michigan State University, 1963.

Lord, Frederick M. The utilization of unreliable difference scores. Journal of Educational Psychology, 1958, 49, 150-152.

Lord, Frederick M. Elementary models for measuring change. Chapter 2 in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Lord, Frederick M., and Novick, Melvin R. Statistical theories of mental test scores. Reading, Mass.: Addison Wesley, 1968.

Magnusson, David. Test theory. Reading, Mass.: Addison Wesley, 1967.

Saupe, Joe L. Technical considerations in measurement. Appendix in Dressel, Paul L. (Ed.) Evaluation in higher education. Boston, Mass.: Houghton Mifflin, 1961.

Saupe, Joe L. Selecting items to measure change. Journal of Educational Measurement, 1966, 3, 223-228.

Shoemaker, David M. Note on the attenuating effect of zero-variance items on K R-20. Journal of Educational Measurement, 1970, 6, 255-256.

Tucker, Ledyard; Damarin, Fred; and Messick, Samuel. A base-free measure of change. Psychometrika, 1966, 31, 457-473.

Webster, Harold, and Bereiter, Carl. The reliability of changes measured by mental test scores. Chapter 3 in Harris, Chester W. (Ed.) Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

APPENDIX

APPENDIX

The 120 items on the Inventory of Beliefs, Form I, used in this study are listed here. Following the listing of items, Tables A.1 and A.2 are presented to indicate the subscales in which each item appeared. Percentages of overlap of items between the scales selected by different item analysis methods are reported in Tables A.3 through A.10.

1. If you want a thing done right, you have to do it yourself.
2. There are times when a father, as head of the family, must tell the other family members what they can and cannot do.
3. Lowering tariffs to admit more foreign goods into this country lowers our standard of living.
4. Literature should not question the basic moral concepts of society.
5. Reviewers and critics of art, music and literature decide what they like and then force their tastes on the public.
6. Why study the past, when there are so many problems of the present to be solved.
7. Business men and manufacturers are more important to society than artists or musicians.
8. There is little chance for a person to advance in business or industry unless he knows the right people.

APPENDIX

The 120 items on the Inventory of Beliefs, Form I, used in this study are listed here. Following the listing of items, Tables A.1 and A.2 are presented to indicate the subscales in which each item appeared. Percentages of overlap of items between the scales selected by different item analysis methods are reported in Tables A.3 through A.10.

1. If you want a thing done right, you have to do it yourself.
2. There are times when a father is head of the family, must tell the other family members what they can and cannot do.
3. Lowering tariffs to admit more foreign goods into this country lowers our standard of living.
4. Literature should not depict the basic social concepts of society.
5. Reviewers and critics of art, music and literature decide what they like and then force their tastes on the public.
6. Why study the past, when there are so many problems of the present to be solved.
7. Business men and manufacturers are more important to society than artists or musicians.
8. There is little chance for a person to advance in business or industry unless he knows the right people.

9. Man has an inherent guide to right and wrong--his conscience.

10. The main thing about good music is lovely melody.

11. It is only natural and right for each person to think that his family is better than any other.

12. All objective data gathered by unbiased persons indicate that the world and universe are without order.

13. Any man can find a job if he really wants to work.

14. We are finding out today that liberals really are soft-headed, gullible, and potentially dangerous.

15. A man can learn as well by striking out on his own as he can by following the advice of others.

16. The predictions of economists about the future of business are no better than guesses.

17. Being a successful wife and mother is more a matter of instinct than of training.

18. A person often has to get mad in order to push others into action.

19. There is only one real standard in judging art works--each to his own taste.

20. Business enterprise, free from government interference, has given us our high standard of living.

21. Nobody can make a million dollars without hurting other people.

22. Anything we do for a good cause is justified.

23. Public resistance to modern art proves that there is something wrong with it.

24. Sending letters and telegrams to congressmen is mostly a waste of time.

25. Many social problems would be solved if we did not have so many immoral and inferior people.

9. Man has an inherent guide to right and wrong--his conscience.

10. The main thing about good music is lovely melody.

11. It is only natural and right for each person to think that his family is better than any other.

12. All objective data gathered by unbiased persons indicate that the world and universe are without order.

13. Any man can find a job if he really wants to work.

14. We are finding out today that liberals really are soft-headed, gullible, and potentially dangerous.

15. A man can learn as well by striking out on his own as he can by following the advice of others.

16. The predictions of economists about the future of business are no better than guesses.

17. Being a successful wife and mother is more a matter of instinct than of training.

18. A person often has to get mad -- really mad -- before he does anything decisive in action.

19. There is only one real satisfaction in judging art work--each to his own taste.

20. Business enterprise, free from government interference, has given us our high standard of living.

21. Nobody can make a million dollars without hurting other people.

22. Anything we do for a good cause is justified.

23. Public resistance to nuclear war proves that there is something wrong with it.

24. Sending letters and telegrams to congressmen is mostly a waste of time.

25. Many social problems would be solved if we did not have so many immoral and inferior people.

26. Art which does not tell a human story is empty.

27. You can't do business on friendship: profits are profits; and good intentions are not evidence in a law court.

28. A person has troubles of his own; he can't afford to worry about other people.

29. Books and movies should start dealing with entertaining or uplifting themes instead of the present unpleasant, immoral, or tragic ones.

30. Children should be made to obey since you have to control them firmly during their formative years.

31. The minds of many youth are being poisoned by bad books.

32. Speak softly, but carry a big stick.

33. Ministers in churches should not preach about economic and political problems.

34. Each man is on his own in life and must determine his own destiny.

35. New machines should be taxed to support the workers they displace.

36. The successful merchant can't allow sentiment to affect his business decisions.

37. Ministers who preach socialistic ideas are a disgrace to the church.

38. Labor unions don't appreciate all the advantages which business and industries have given them.

39. It's only natural that a person should take advantage of every opportunity to promote his own welfare.

40. We should impose a strong censorship on the morality of books and movies.

41. The poor will always be with us.

26. Art which does not tell a human story is empty.

27. You can't do business on friendship: profits are profits; and good intentions are not evidence in a law court.

28. A person has troubles of his own; he can't afford to worry about other people.

29. Books and movies should start dealing with entertaining or uplifting themes instead of the present unpleasant, immoral, or tragic ones.

30. Children should be made to obey since you have to control them firmly during their formative years.

31. The minds of many youth are being poisoned by bad books.

32. Speak softly, but carry a big stick.

33. Ministers in churches should not preach about economic and political problems.

34. Each man is on his own in life and must determine his own destiny.

35. New machines should be used to supplant the workers they displace.

36. The successful [illegible] all the resentment to affect his business decisions.

37. Minister who preach on [illegible] bring disgrace to the church.

38. Labor unions don't appreciate all the advantages which business and industries have given them.

39. It's only natural that a person should take advantage of every opportunity to promote his own welfare.

40. We should impose a strong censorship on the morality of books and movies.

41. The poor will always be with us.

42. A person who is incapable of real anger must also be lacking in moral conviction.

43. If we allow more immigrants into this country, we will lower our standard of culture.

44. People who live in the slums have no sense of respectability.

45. We acquire the highest form of freedom when our wishes conform to the will of society.

46. Modern paintings look like something dreamed up in a horrible nightmare.

47. Voting determines whether or not a country is democratic.

48. The government is more interested in winning elections than in the welfare of the people.

49. Feeble-minded people should be sterilized.

50. In our society, a person's first duty is to protect from harm himself and those dear to him.

51. Those who can, do; those who can't, teach.

52. The best government is one which governs least.

53. History shows that every great nation was destroyed when its people became soft and its morals lax.

54. Philosophers on the whole act as if they were superior to ordinary people.

55. A woman who is a wife and mother should not try to work outside the home.

56. We would be better off if people would talk less and work more.

57. In some elections there is not much point in voting because the outcome is fairly certain.

58. The old masters were the only artists who <u>really</u> knew how to draw and paint.

42. A person who is incapable of real anger must also be lacking in moral conviction.

43. If we allow more immigrants into this country, we will lower our standard of culture.

44. People who live in the slums have no sense of respectability.

45. We acquire the highest form of freedom when our wishes conform to the will of society.

46. Modern paintings look like something dreamed up in a horrible nightmare.

47. Voting determines whether or not a country is democratic.

48. The government is more interested in winning elections than in the welfare of the people.

49. Feeble-minded people should be sterilized.

50. In our society, a person's first duty is to protect from harm himself and those dear to him.

51. Those who can, do; those who can't, teach.

52. The best government is one which governs least.

53. History shows that every [illegible] was destroyed when its people became soft and [illegible]

54. Philosophers on the [illegible] if they were superior to ordinary people.

55. A woman who is a wife and [illegible] should not try to work outside the home.

56. We would be better off if people would talk less and work more.

57. In some elections there is not much point in voting because the outcome is fairly certain

58. The old masters were the only artists who really knew how to draw and paint

59. Most intellectuals would be lost if they had to make a living in the realistic world of business.

60. You cannot lead a truly happy life without strong moral and religious convictions.

61. If we didn't have strict immigration laws, our country would be flooded with foreigners.

62. When things seem black, a person should not complain, for it may be God's will.

63. Miracles have always taken place whenever the need for them has been great enough.

64. Science is infringing upon religion when it attempts to delve into the origin of life itself.

65. A person has to stand up for his rights or people will take advantage of him.

66. A lot of teachers, these days, have radical ideas which need to be carefully watched.

67. Now that America is the leading country in the world, it's only natural that other countries should try to be like us.

68. Most Negroes would become overbearing and disagreeable if not kept in their place.

69. Foreign films emphasize sex more than American films do.

70. Our rising divorce rate is a sign that we should return to the values which our grandparents held.

71. Army training will be good for most modern youth because of the strict discipline they will get.

72. When operas are sung in this country they ought to be translated into English.

73. People who say they're religious but don't go to church are just hypocrites.

74. What the country needs, more than laws or politics, is a few fearless and devoted leaders in whom the people can have faith.

75. Pride in craftsmanship and in doing an honest day's work is a rare thing these days.

76. The United States may not have had much experience in international dealings but it is the only nation to which the world can turn for leadership.

77. In practical situations, theory is of very little help.

78. No task is too great or too difficult when we know that God is on our side.

79. A sexual pervert is an insult to humanity and should be punished severely.

80. A lot of science is just using big words to describe things which many people already know through common sense.

81. Manual labor and unskilled jobs seem to fit the Negro mentality and ability better than more skilled or responsible work.

82. A person gets what's coming to him in this life if he doesn't believe in God.

83. Public officials may try to be honest but they are caught in a web of influence which tends to corrupt them.

84. Science makes progress only when it attempts to solve urgent practical problems.

85. Most things in life are governed by forces over which we have no control.

86. Young people today are in general more immoral and irresponsible than young people of previous generations.

87. Americans may tend to be materialistic, but at least they aren't cynical and decadent like most Europeans.

88. The many different kinds of children in school these days force teachers to make a lot of rules and regulations so that things will run smoothly.

14. What the country needs, more than laws or politics, is a few fearless and devoted leaders in whom the people can have faith.

15. Pride in craftsmanship and in doing an honest day's work is a rare thing these days.

16. The United States may not have had much experience in international dealings but it is the only nation to which the world can turn for leadership.

17. In practical situations, theory is of very little help.

18. No task is too great or too difficult when we know that God is on our side.

19. A sexual pervert is an insult to humanity and should be punished severely.

20. A lot of science is just using big words to describe things which many people already know through common sense.

21. Manual labor and unskilled jobs seem to be the keys to mentality and ability better than more skilled or responsible work.

22. A person gets what's coming to him at this time if he doesn't believe in God.

23. Public officials may use [illegible] but they are caught in a web of influence which binds [illegible] them.

24. Science makes progress only when it attempts to solve urgent practical problems.

25. Most things in life are governed by forces over which we have no control.

26. Young people today are in general more mature and irresponsible than young people of previous generations.

27. Americans may tend to be materialistic, but at least they aren't cynical and decadent like most Europeans.

28. The many different kinds of clothing in recent times days force teachers to make a lot of rules and regulations so that things will run smoothly.

89. Jews will marry out of their own religious group whenever they have the chance.

90. The worst danger to real Americanism during the last 50 years has come from foreign ideas and agitators.

91. Europeans criticise the United States for its materialism but such criticism is only to cover up their realization that American culture is far superior to their own.

92. The scientist that really counts is the one who turns theories into practical use.

93. No one can really feel safe when scientists continue to explore whatever they wish without any social or moral restraint.

94. Nudist colonies are a threat to the moral life of a nation.

95. One trouble with Jewish businessmen is that they stick together and prevent other people from having a fair chance in competition.

96. No world organization should have the right to tell Americans what they can or cannot do.

97. There is a source of knowledge that is not dependent upon observation.

98. Despite the material advantages of today, family life now is not as wholesome as it used to be.

99. The United States doesn't have to depend on the rest of the world in order to be strong and self-sufficient.

100. Foreigners usually have peculiar and annoying habits.

101. Parents know as much about how to teach children as public school teachers.

102. The best assurance of peace is for the United States to have the strongest army, navy, air force, and the most atom bombs.

103. Some day machinery will do nearly all of man's work, and we can live in leisure.

89. Jews will marry out of their own religious group whenever they have the chance.

90. The worst danger to real Americanism during the last 50 years has come from foreign ideas and agitators.

91. Europeans criticize the United States for its materialism but such criticism is only to cover up their realization that American culture is far superior to their own.

92. The scientist that really counts is the one who turns theories into practical use.

93. No one can really feel safe when scientists continue to explore whatever they wish without any social or moral restraint.

94. Nudist colonies are a threat to the moral life of a nation.

95. One trouble with Jewish businessmen is that they stick together and prevent other people from having a fair chance in competition.

96. No world organization should be formed right to tell Americans what they can or cannot do.

97. There is a source of knowledge that is not dependent upon observation.

98. Despite the material [illegible] family life now is not as wholesome as it used to be.

99. The United States must [illegible] in order to be strong and [illegible].

100. Foreigners usually have peculiar and annoying habits.

101. Parents know as much about how to teach children as public school teachers.

102. The best assurance of peace is for the United States to have the strongest army, navy, air force, and the most atom bombs.

103. Some day machinery will do nearly all of man's work, and we can live in leisure.

104. There are too many people in this world who do nothing but think about the opposite sex.

105. Modern people are superficial and tend to lack the finer qualities of manhood and womanhood.

106. Members of religious sects who refuse to salute the flag should be punished for their lack of patriotism.

107. Political parties are run by insiders who are not concerned with the public welfare.

108. As young people grow up they ought to get over their radical ideas.

109. Negroes have their rights, but it is best to keep them in their own districts and schools and to prevent too much contact with whites.

110. The twentieth century has not had leaders with the vision and capacity of the founders of this country.

111. There are a lot of things in this world that will never be explained by science.

112. Sexual relations between brother and sister are contrary to natural law.

113. There may be a few exceptions, but in general Jews are pretty much alike.

114. The world will get so bad that some of these times God will destroy it.

115. Children should learn to respect and obey their teachers.

116. Other countries don' t appreciate as much as they should all the help that America has given them.

117. We would be better off if there were fewer psychoanalysts probing and delving into the human mind.

118. American free enterprise is the greatest bulwark of democracy.

104. There are too many people in this world who do nothing but think about the opposite sex.

105. Modern people are superficial and tend to lack the finer qualities of manhood and womanhood.

106. Members of religious sects who refuse to salute the flag should be punished for their lack of patriotism.

107. Political parties are run by insiders who are not concerned with the public welfare.

108. As young people grow up they ought to get over their radical ideas.

109. Negroes have their rights, but it is best to keep them in their own districts and schools and to prevent too much contact with whites.

110. The twentieth century has not had leaders with the vision and capacity of the founders of this country.

111. There are a lot of things in this world that will never be explained by science.

112. Sexual relations here [illegible] are contrary to natural law.

113. There may be [illegible] laws are pretty much alike.

114. The world will get to [illegible] destroy it.

115. Children should learn to respect [illegible].

116. Other countries don't [illegible] help that America has given them.

117. We would be [illegible] probing and defying into the human mind.

118. American free enterprise is the greatest bulwark of democracy.

119. If a person is honest, works hard, and trusts in God, he will reap material as well as spiritual rewards.

120. One will learn more in the school of hard knocks than he ever can from a textbook.

TABLE A. 1. --Listing of subscales in which each change-item first appeared after item analysis with one-to-four scoring system.*

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Saupe r	IV Random
1	--	60	90	60
2	--	15	--	15
3	15	60	90	90
4	30	90	--	30
5	15	90	30	--
6	30	--	90	90
7	--	--	60	60
8	--	--	--	30
9	30	30	--	15
10	90	60	60	--
11	60	60	60	90
12	--	--	--	30
13	90	15	--	60
14	90	--	30	90
15	60	60	60	60
16	--	90	--	30
17	60	90	--	15
18	60	90	--	60

*The numbers in the table indicate the scale length when the item first appeared. If an item is included in a scale of 15 items, it is obviously included in all scales using the same procedure which are of greater length.

TABLE A.1.--Listing of subscales in which each change-item first appeared after item analysis with one-to-four scoring system.*

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Sample r	IV Random
1	--	60	90	60
2	--	15	--	15
3	15	60	90	90
4	30	90	--	30
5	15	90	90	--
6	30	--	90	60
7	--	-	60	60
8	--		--	30
9	30	90		15
10	90	60	60	--
11	60	90	90	90
12	--	--		30
13	30	15		60
14	60		30	90
15	30	90	30	60
16	--	90	--	30
17	30	90	--	15
18	60	90	--	60

*The numbers in the table indicate the scale length when the item first appeared. If an item is included in a scale of 15 items, it is obviously included in all scales using the same procedure which are of greater length.

TABLE A.1. --Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Saupe r	IV Random
19	30	15	30	90
20	30	60	30	--
21	60	90	90	--
22	60	90	--	90
23	--	--	90	60
24	60	--	90	30
25	15	60	15	--
26	60	90	--	60
27	15	60	60	90
28	--	--	90	90
29	30	30	60	--
30	60	15	90	90
31	30	30	60	60
32	30	60	--	30
33	15	90	--	15
34	15	15	--	60
35	90	--	30	--
36	60	90	90	90
37	90	90	30	--
38	15	15	--	90
39	--	30	60	60
40	30	60	90	30
41	--	15	--	15

TABLE A.1. --Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Slope r	IV Random
19	30	15	30	90
20	30	60	30	--
21	60	90	90	--
22	60	90	--	30
23	--	--	90	60
24	60	--	90	30
25	15	60	15	--
26	60	90	--	60
27	15	60	60	90
28	--	--	90	60
29	30	60	60	--
30	60	90	90	90
31	30	60	60	90
32	30	90	--	30
33	15	90	--	15
34	15	60	--	60
35	90	--	30	--
36	60	90	90	90
37	90	90	30	--
38	15	15	--	90
39	--	30	60	90
40	30	60	90	30
41	--	15	--	15

TABLE A.1. --Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Saupe r	IV Random
42	60	90	--	60
43	90	--	60	--
44	--	--	90	90
45	15	60	30	90
46	30	60	90	60
47	15	90	90	60
48	60	--	--	30
49	60	60	--	15
50	--	15	60	60
51	30	--	--	15
52	15	90	--	90
53	90	15	30	90
54	90	60	60	--
55	60	15	60	60
56	90	30	60	30
57	--	--	90	60
58	--	--	90	90
59	90	90	90	90
60	30	30	--	15
61	90	30	90	--
62	90	30	60	90
63	90	60	15	60
64	30	90	60	30

TABLE A. 1. --Continued.

Item Number	Item Analysis Method			
	I Chegge Variance	II Pretest	III Sange r	IV Random
42	60	90	--	60
43	90	--	60	--
44	--	--	90	90
45	15	60	30	30
46	30	30	90	60
47	15	90	90	60
48	60	--	--	30
49	60	60	--	15
50	--	15	30	30
51	30	--	--	15
52	15	30	--	90
53	90	--	30	60
54	90	60	60	--
55	60	15	60	90
56	90	30	60	30
57	--	--	90	30
58	--	--	90	90
59	90	90	30	30
60	30	30	--	15
61	45	30	30	--
62	90	45	60	90
63	90	30	15	60
64	30	90	60	30

TABLE A. 1. --Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Saupe r	IV Random
65	--	30	--	15
66	60	90	60	60
67	90	60	30	--
68	90	--	90	90
69	15	30	15	90
70	90	60	60	--
71	60	15	60	60
72	90	60	60	30
73	30	90	15	15
74	60	90	--	60
75	60	60	90	90
76	60	60	60	--
77	--	90	30	90
78	60	15	30	--
79	60	60	90	60
80	90	--	15	--
81	90	--	30	15
82	60	90	60	60
83	90	60	90	--
84	--	--	60	--
85	60	--	--	60
86	60	--	60	--
87	--	90	15	60

TABLE A. 1. --Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Shape r	IV Random
85	--	30	--	15
86	60	90	50	80
87	90	80	30	--
88	30	--	90	90
89	15	30	15	90
70	90	80	80	--
71	60	15	90	80
72	30	60	60	30
73	30	90	15	15
74	60	90	--	90
75	60	60	90	90
76	60	60	60	--
77	--	--	90	90
78	90	--	75	--
79	90	--	--	90
80	90	--	15	--
81	90	--	30	15
82	60	90	60	30
83	90	90	90	--
84	--	--	30	--
85	90	--	--	60
86	90	--	60	--
87	--	90	15	90

TABLE A. 1. --Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Saupe r	IV Random
88	90	30	15	--
89	--	--	90	15
90	60	60	--	60
91	90	60	15	--
92	15	15	15	60
93	--	90	15	--
94	--	60	90	60
95	90	90	15	--
96	60	90	90	30
97	60	60	--	15
98	--	60	60	60
99	--	--	90	--
100	--	--	90	--
101	90	--	--	60
102	15	--	15	90
103	60	90	90	60
104	--	60	60	30
105	30	--	30	15
106	--	90	30	90
107	--	90	60	--
108	--	60	90	--
109	--	--	15	90
110	90	--	90	90

TABLE A. 1.--Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Serge r	IV Random
88	90	30	15	--
89	--	--	90	15
90	60	60	--	60
91	90	60	15	--
92	15	15	15	60
93	--	90	15	--
94	--	60	30	60
95	30	90	15	--
96	60	60	90	30
97	60	60		15
98	--	60	60	60
99	--			--
100	--		60	--
101	90			60
102	15		15	90
103	60	90	90	60
104	--	90	90	30
105	30		30	15
106	--	90	30	90
107	--	60	60	
108	--	60	90	--
109	--		15	90
110	60	-	90	90

TABLE A. 1. --Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Saupe r	IV Random
111	90	30	--	30
112	15	30	--	15
113	60	90	15	--
114	90	90	15	--
115	--	15	90	90
116	90	30	30	--
117	90	--	60	90
118	60	30	90	90
119	90	15	60	30
120	15	60	60	60

TABLE A. 1.--Continued.

Item Number	Item Analysis Method			
	I Change Variance	II Pretest	III Surge r	IV Random
111	90	30	--	30
112	15	30	--	15
113	60	60	15	--
114	90	90	15	--
115	--	15	90	90
116	90	30	30	--
117	90	--	60	90
118	90	30	90	90
119	90	15	45	30
120	15	45	45	60

TABLE A.2. --Listing of subscales in which each change-item first appeared after item analysis with zero-one scoring.*

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Saupe r	IV Triserial r	V Random
1	60	60	90	90	60
2	--	15	--	--	15
3	15	60	--	--	90
4	30	--	15	60	30
5	15	90	90	15	--
6	--	--	90	60	90
7	90	--	15	15	60
8	90	--	--	--	30
9	60	15	--	--	15
10	30	60	60	60	--
11	30	60	--	--	90
12	--	--	--	--	30
13	--	15	90	90	60
14	--	--	90	90	90
15	30	90	--	--	60
16	--	--	--	--	30
17	30	90	90	--	15
18	60	60	--	--	60

*The numbers in the table indicate the scale length when the item first appeared. If an item is included in a scale of 15 items, it is obviously included in all scales using the same procedure which are of greater length.

TABLE A.2. -- Listing of subscales in which each change-item first appeared after item analysis with zero-one scoring. *

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Scope r	IV Tetrachoric r	V Random
1	60	60	90	90	60
2	--	15	--	--	15
3	15	60	--	--	60
4	30	--	15	60	30
5	15	90	90	15	--
6	--	--	60	60	90
7	90	--	15	15	60
8	90	--	--	--	30
9	60	15	--	--	15
10	30	60	60	60	--
11	30	45	--	--	30
12	--	--	--	--	90
13	--	15	--	90	60
14	--	--	30	--	90
15	30	90	--	--	60
16	--	--	--	--	30
17	30	--	90	--	15
18	60	60	--	--	60

*The numbers in the table indicate the scale length when the item first appeared. If an item is included in a scale of 15 items, it is obviously included in all scales using the same procedure which are of greater length.

TABLE A.2. --Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Saupe r	IV Triserial r	V Random
19	60	30	30	30	90
20	30	30	90	90	--
21	30	90	--	--	--
22	60	90	90	90	90
23	--	--	--	--	60
24	90	--	--	--	30
25	30	60	30	30	--
26	90	90	--	--	60
27	15	60	60	90	90
28	--	--	90	90	90
29	15	60	15	15	--
30	--	15	90	90	90
31	60	30	--	30	60
32	15	60	60	90	30
33	30	90	--	--	15
34	60	15	--	--	60
35	90	--	30	30	--
36	15	60	90	90	90
37	60	90	--	90	--
38	15	30	60	60	90
39	90	15	60	60	60
40	60	60	90	90	30

TABLE A.3.--Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Scope r	IV Trisserial r	V Random
19	60	30	30	30	90
20	30	30	90	60	--
21	30	90	--	--	--
22	60	90	90	90	90
23	--	--	--	--	60
24	90	--	--	--	30
25	30	60	30	30	--
26	90	90	--	--	60
27	15	60	60	90	90
28	--	--	90	90	90
29	15	60	15	15	--
30	--	--	60	60	60
31	60	15	--	15	60
32	15	60	60	60	30
33	30	60	--	--	15
34	60	15	--	--	60
35	90	--	60	30	--
36	15	60	60	60	90
37	60	60	--	90	--
38	15	60	60	60	90
39	90	15	60	60	60
40	60	60	30	90	30

TABLE A.2. --Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Saupe r	IV Triserial r	V Random
41	--	15	90	60	15
42	90	90	--	--	60
43	--	--	--	--	--
44	90	--	60	60	90
45	15	60	30	30	90
46	15	60	60	60	60
47	60	90	90	90	60
48	90	--	--	--	30
49	90	90	90	90	15
50	--	30	--	--	60
51	90	--	30	60	15
52	15	90	60	90	90
53	90	30	60	60	90
54	30	90	60	60	--
55	30	30	60	60	60
56	60	60	60	60	30
57	--	--	90	60	60
58	--	--	60	30	90
59	15	90	60	90	90
60	60	30	--	--	15
61	--	15	60	60	--
62	90	30	60	90	90

TABLE A.2.--Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Sange r	IV Trimerial r	V Random
41	--	15	30	30	15
42	30	30	--	--	60
43	--	--	--	--	--
44	90	--	60	60	90
45	15	60	30	30	30
46	15	60	60	60	60
47	60	30	30	30	60
48	60	--			30
49	90	90		90	15
50	--	30			60
51	90	--	30	30	15
52	15	75	75	90	90
53	90	30		60	30
54	30	--		90	--
55	30	30	90	90	60
56	60	90	60	60	90
57	--		30	30	60
58	--	--	30	30	30
59	15	90	90	90	30
60	60	30	--	--	15
61	--	15	60	60	--
62	90	30	90	90	90

TABLE A.2. --Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Saupe r	IV Triserial r	V Random
63	60	60	15	15	60
64	90	90	60	60	30
65	--	15	60	60	15
66	60	90	60	60	60
67	90	60	30	60	--
68	--	90	90	90	90
69	15	30	15	15	90
70	60	60	90	90	--
71	90	15	90	90	60
72	--	60	30	15	30
73	90	90	15	15	15
74	90	90	--	--	60
75	60	90	60	60	90
76	60	60	90	90	--
77	90	--	60	60	90
78	60	15	15	15	--
79	60	60	90	90	60
80	--	--	30	30	--
81	--	90	30	30	15
82	90	90	--	--	60
83	15	60	15	15	--
84	90	--	60	60	--

TABLE A.3.--Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Snape r	IV Triserial r	V Random
62	60	60	15	15	60
64	30	30	60	60	30
65	--	15	00	00	15
66	00	00	00	00	60
67	90	60	30	60	--
68	--	90	90	90	90
69	15	30	15	15	90
70	60	60	00	90	--
71	90	15	00	90	90
72	--	10	30	15	30
73	90	90	15	15	15
74	90	90	15	15	90
75	60	90	90	60	30
76	60	60	60	90	--
77	90	15	15	90	90
78	60	15	15	15	--
79	60	60	90	90	60
80	--	00	30	30	--
81	--	90	60	30	15
82	90	90	--	--	60
83	15	60	15	15	--
84	30	--	90	90	--

TABLE A.2. --Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Saupe r	IV Triserial r	V Random
85	60	60	60	90	60
86	60	90	--	--	--
87	60	90	30	30	60
88	60	30	15	15	--
89	--	--	--	--	15
90	60	60	60	60	60
91	30	60	15	15	--
92	30	30	15	15	60
93	--	--	30	30	--
94	90	60	--	--	60
95	90	90	60	60	--
96	90	90	90	--	30
97	60	30	--	--	15
98	--	60	30	30	60
99	--	--	90	60	--
100	--	--	15	15	--
101	--	--	90	90	60
102	60	90	15	15	90
103	60	60	--	--	60
104	90	60	60	30	30
105	30	90	15	30	15
106	--	--	90	90	90

TABLE A.2.--Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Supe r	IV Trimestral r	V Random
85	60	60	60	60	60
86	60	90	--	--	--
87	60	90	30	30	60
88	60	30	15	15	--
89	--	--	--	--	15
90	60	60	60	60	60
91	30	60	15	15	--
92	30	30	15	15	60
93	--	--	60	30	--
94	90	90	--	--	60
95	90	90	60	60	--
96	90	90	90	--	30
97	60	90	--	--	15
98	--	90	90	90	60
99	--	--	90	90	--
100	--	--	15	15	--
101	--	--	90	90	60
102	60	60	15	15	30
103	60	90	--	--	60
104	90	60	60	30	30
105	30	60	15	30	15
106	--	--	90	90	90

TABLE A.2. --Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Saupe r	IV Triserial r	V Random
107	--	--	90	90	--
108	15	90	60	90	--
109	--	--	90	60	90
110	90	--	90	90	90
111	90	30	90	90	30
112	60	15	--	--	15
113	60	60	15	15	--
114	90	90	30	60	--
115	90	15	60	30	90
116	60	30	30	60	--
117	90	--	60	60	90
118	30	15	60	60	90
119	--	15	90	90	30
120	15	60	30	30	60

TABLE A.2.--Continued.

Item Number	Item Analysis Method				
	I Change Variance	II Pretest	III Range r	IV Triserial r	V Random
107	--	--	80	60	--
108	15	60	60	60	--
109	--	--	80	60	60
110	60	--	60	30	60
111	80	30	60	60	30
112	60	15	--	--	15
113	60	60	15	15	--
114	30	60	30	60	--
115	30	15	60	30	60
116	60	30	30	60	--
117	30	--	60	60	60
118	30	15	60	60	60
119	--	15	60	30	30
120	15	60	60	30	60

TABLE A. 3. --Percentage of item overlap for scales chosen by different item analysis methods--15 items, one-to-four scoring.

	Change Variance	Pretest	Saupe r
Change Variance		32	27
Pretest Frequency			7

TABLE A. 4. --Percentage of item overlap for scales chosen by different item analysis methods--30 items, one-to-four scoring.

	Change Variance	Pretest	Saupe r
Change Variance		33	33
Pretest Frequency			20

TABLE A. 5. --Percentage of item overlap for scales chosen by different item analysis methods--60 items, one-to-four scoring.

	Change Variance	Pretest	Saupe r
Change Variance		80	73
Pretest Frequency			73

TABLE A.3. --Percentage of item overlap for scales chosen by different item analysis methods--15 items, one-to-four scoring.

	Change Variance	Pretest	Range r
Change Variance		33	27
Pretest Frequency			7

TABLE A.4. --Percentage of item overlap for scales chosen by different item analysis methods--30 items, one-to-four scoring.

	Change Variance	Pretest	Range r
Change Variance		43	43
Pretest Frequency			20

TABLE A.5. --Percentage of item overlap for scales chosen by different item analysis methods--60 items, one-to-four scoring.

	Change Variance	Pretest	Range r
Change Variance		60	73
Pretest Frequency			73

TABLE A.6. --Percentage of item overlap for scales chosen by different item analysis methods--90 items, one-to-four scoring.

	Change Variance	Pretest	Saupe r
Change Variance		80	73
Pretest Frequency			73

TABLE A.7. --Percentage of item overlap for scales chosen by different item analysis methods--15 items, zero-one scoring.

	Change Variance	Pretest	Saupe r	Triserial r
Change Variance		0	20	20
Pretest Frequency			7	7
Saupe r				87
Triserial r				

TABLE A.6.--Percentage of item overlap for scales chosen by different item analysis methods--80 items, one-to-four scoring.

	Change Variance	Protest	Range r
Change Variance		80	73
Protest Frequency			73

TABLE A.7.--Percentage of item overlap for scales chosen by different item analysis methods--75 items, zero-one scoring.

	Change Variance	Protest	Range r	Triserial r
Change Variance		0	20	20
Protest Frequency			7	7
Range r				87
Triserial r				

TABLE A. 8. --Percentage of item overlap for scales chosen by different item analysis methods--30 items, zero-one scoring.

	Change Variance	Pretest	Saupe r	Triserial r
Change Variance		27	33	33
Pretest Frequency			20	23
Saupe r				83
Triserial r				

TABLE A. 9. --Percentage of item overlap for scales chosen by different item analysis methods--60 items, zero-one scoring.

	Change Variance	Pretest	Saupe r	Triserial r
Change Variance		68	57	50
Pretest Frequency			57	53
Saupe r				89
Triserial r				

TABLE A.8. --Percentage of item overlap for scales chosen by different item analysis methods--30 items. zero-one scoring.

	Change Variance	Pretest	Shape r	Triserial r
Change Variance		27	38	32
Pretest Frequency			20	23
Shape r				83
Triserial r				

TABLE A.9. --Percentage of item overlap for scales chosen by different item analysis methods--60 items. zero-one scoring

	Change Variance	Pretest	Shape r	Triserial r
Change Variance		[illegible]	[illegible]	50
Pretest Frequency			[illegible]	43
Shape r				84
Triserial r				

TABLE A. 10. --Percentage of item overlap for scales chosen by different item analysis methods--90 items, zero-one scoring.

	Change Variance	Pretest	Saupe r	Triserial r
Change Variance		87	72	73
Pretest Frequency			76	76
Saupe r				98
Triserial r				

TABLE A. 19. --Percentage of item overlap for scales chosen by different item analysis methods--30 items, zero-one scoring.

	Change Variance	Pretest	Sauge r.	Trisserial r
Change Variance		87	72	73
Pretest Frequency			76	78
Sauge r				98
Trisserial r				