AN ITEM ANALYSIS TECHNIQUE BASED UPON ADJACENT GROUP DIFFERENCES

Thesis for the Degree of Ed. D.

MICHIGAN STATE COLLEGE

Robert Arthur Jackson

1952

This is to certify that the

thesis entitled

An Item-Analysis Technique

Based Upon Adjacent-Group Differences

presented by

Robert Arthur Jackson

has been accepted towards fulfillment
of the requirements for

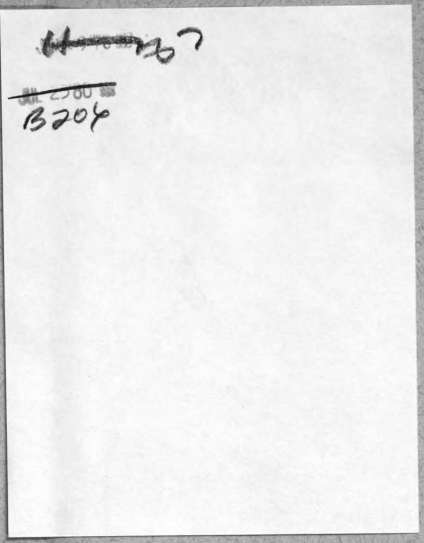
Ed. D. degree in Education

H. W. Sundwall

H. W. Sundwall
Major professor

Date December 2, 1952

O-169


B204

AN ITEM ANALYSIS TECHNIQUE BASED UPON ADJACENT GROUP DIFFERENCES

By

Robert Arthur Jackson

A THESIS

Submitted to the School of Graduate Studies of Michigan State College of Agriculture and Applied Science in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

Department of Education

1952

ACKNOWLEDGMENTS

To Dr. H. W. Sundwall, the author is very grateful for the patience, encouragement, and direction during the period of dissertation-preparation.

To Dr. W. D. Baten is due gratitude for criticism of the statistical portion of the manuscript.

To Dr. M. Muntyan, the author is deeply indebted for his critical examination of the manuscript.

To Dr. V. H. Noll, the author wishes to extend his sincere thanks for his guidance and supervision in the period he directed the Guidance Committee.

AN ITEM ANALYSIS TECHNIQUE BASED UPON ADJACENT GROUP DIFFERENCES

By

Robert Arthur Jackson

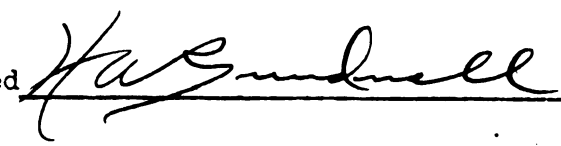
AN ABSTRACT

Submitted to the School of Graduate Studies of Michigan State College of Agriculture and Applied Science in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

Department of Education

1952

Approved 

The major purpose of this study was to present and evaluate a new item-analysis technique applicable in situations where the primary interest is in the discrimination between the members of two or more groups, rather than discriminating between the members within groups. This type of problem occurs frequently in assigning letter grades and in selection work where individuals are to be divided into two groups (that is, those who may be expected to succeed and those who may be expected to fail in a particular situation). A procedure for computing the adjacent-group item-validity indices was presented. This adjacent-group technique resulted in a maximized ratio of between-groups variance to total variance.

It was assumed in this investigation that the test-score distribution should be the one best fitting the need of the particular situation where the test is to be used. In the case of two groups, the most discriminating test was found to be one that yields a score distribution with a point of partition at the abscissa of the minimal ordinate between the two group modes. A theoretical examination of the score distribution for two groups showed that a non-overlapping bimodal distribution may be obtained by selecting a sufficient number of appropriate items. In the theoretical comparison of the bimodal test-score distribution with a normal test-score distribution, it was demonstrated that the bimodal distribution resulted in fewer chance errors than the normal distribution. In the case of more than two groups, the distribution should have points of partition at the abscissa of the minimal ordinate between any two adjacent-group modes.

The empirical findings of this study indicated that the selected test items tended to be stable under cross-validation. The empirical studies on the error of measurement of the multimodal test-score distribution also showed this error to be minimal at the points of partition separating two adjacent groups; the tendency seemed to be that the error of measurement approaches zero at the points of partition. These findings were interpreted in terms of the small number of cases involved in the data.

Since test items discriminating perfectly between two adjacent groups are difficult to obtain, it is quite apparent that the adjacent-group technique for the selection of items is used to greater advantage in situations where a large source of test items is available. However, the adjacent-group technique can also be applied in situations where intra-group comparisons are to be made and the source of items is more limited. This technique was found to be as satisfactory as Horst's more laborious technique of maximizing function in the selection of the most valid items in terms of an external criterion. It was found to be superior to the technique of Flanagan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

In the construction of a test for measurement purposes, the test writer is confronted with a two-fold problem. He must adjust the length of the test to stay within the amount of time available and to avoid fatiguing the individuals taking the test. At the same time, he must make certain that the items in the test constitute an adequate sampling of all possible items related to the trait being measured. A basic purpose of a test is to place individuals along a scale for measurement of a given trait in accordance with real differences. This means that a test used in the measurement of a trait must possess discriminative power; and since tests are made up of individual items, each item should contribute to this discrimination. The original construction of items, which are to represent the theoretical pool of possible items, depends upon the skill and judgment of the test writer. Since the personal judgment of an individual is subject to error, many statistical processes, called item analysis techniques, have been utilized to evaluate each of the test items. All item analysis techniques are subject to certain limitations:

(1) no item analysis technique can by itself turn poor items into good items or operate satisfactorily without a reliable criterion;

(2) the results obtained by item analysis techniques must be understood before they may be used efficiently;

(3) item analysis results from one experimental group may not be exactly parallel for another group; and

(4) the item analysis data should supplement, not supplant, subjective opinion.

The test items are classified as satisfactory or unsatisfactory by examining two statistical characteristics for each item; (1) the difficulty of each item (percent of the students failing to succeed on the item); and (2) an index of discrimination (degree to which the item is effective in differentiating between those who are high and those who are low in respect to the trait being measured). A satisfactory item would not be failed or passed by all of the students; it would be passed by students who possess the trait to a high degree more often than students who possess the trait to a low degree. An item which was not satisfactory would be passed by the lower individual more often than the higher one. Unsatisfactory test items occur when the item and the general criterion are not measuring the same trait. Satisfactory test items should function well; they should have a firm theoretical basis.

It seems desirable to point out that a fundamental assumption, underlying all item analysis techniques, is that the items differ from one to another in respect to difficulty and discrimination. Merrill stated:

> If the items are heterogeneous with respect to validity, one can say with some confidence that the most valid items in one sample will in general be the most valid in any other sample, and the use of good items for predictive purposes is therefore justified. In the event, however, that there is a strong probability of the items being homogeneous, there is no justification for any selection.[1]

A great variety of procedures have been employed to determine which items should be selected for a test. These procedures yield statistical data to be used as a guide in assembling the final form of the test, and they do not take the place of ability in item construction. Useful sur-

---

[1]W. W. Merrill, "Sampling Theory in Item Analysis," Psychometrika, II, pp. 215-6, 1937.

veys of indices of validity or consistency have been provided by Lentz, et. al.,[1] Lindquist and Cook[2], Zubin[3], Long and Sandiford[4], Guilford[5], Swineford[6], and Davis[7].

When item selection techniques are applied, two major types of situations are encountered. In the first type, we are relating the performance of the individuals on an item to their performance on some type of continuous measure. This continuous measure is usually the total score on the test, but it may be some external criterion. The second type is one in which the individuals' performance on the item is related to a dichotomous grouping on the criterion variable.

Item Analysis With a Continuous Score

When the variable with which the item is being analyzed is continuously distributed, two statistical approaches are possible. In one, the

---

[1]T. F. Lentz, B. Hirshstein, and J. H. Finch, "Evaluation of Methods of Evaluating Test Items," Journal of Educational Psychology, XXIII, 344-350, 1932.

[2]E. F. Lindquist and W. W. Cook, "Experimental Procedures in Test Evaluation," Journal of Experimental Education, I, 163-185, 1933.

[3]J. Zubin, "The Method of Internal Consistency for Selecting Test Items," Journal of Educational Psychology, XXI, 345-356, 1934.

[4]J. A. Long and P. Sandiford, The Validation of Test Items. Toronto: Department of Educational Research, University of Toronto, Bulletin No. 3, 1935, pp. 126.

[5]J. P. Guilford, Psychometric Methods. New York: McGraw-Hill, pp. 428-437, 1936.

[6]F. Swineford, "Validity of Test Items," Journal of Educational Psychology, XXIVV, 68-78, 1936.

[7]F. B. Davis, Chapter 9. Item Selection Techniques. E. F. Lindquist, et. al., Educational Measurement. Washington: American Council on Education, pp. 266-328, 1951.

degree of relationship between success on the item and the criterion score is determined by using r_{bis}, r_p, η, and V_{mlb}.[1] All of these indices are dependent upon the proportion of the group answering the item correctly, the standard deviation of the criterion scores of the entire group, and the difference between the mean criterion score of the students answering the item correctly and the mean criterion score of the students answering the item incorrectly. The second approach is dependent upon the difference in the criterion scores of the individuals passing the item and those failing. Two statistical techniques used to indicate whether there is a significant relationship between the performance of the students on the item and the criterion are the standardized difference between the means and the F-ratio.

The simple difference between the mean criterion score of the individuals answering the item right and those answering it wrong, or the overlapping methods derived from the proportion of the individuals failing the item whose criterion scores exceeded the median scores of those passing the item yield rough indications of difference.[2]

Item Analysis With Groups

Some simplified item analysis procedures have been developed for use when the criterion scores are treated as a dichotomy. When the criterion variable is a natural dichotomy these techniques must be used; these procedures may also be used when a continuous criterion is divided into a

---

[1] Long and Sandiford, op. cit., pp. 24-29.

[2] Ibid.

dichotomy for ease of computation. In treating a continuous variable as a dichotomy, an arbitrary dividing line is set up for the continuous score; those individuals falling below the dividing line constitute one group and those with scores greater than the dividing score constitute the other. Most of the techniques used with a dichotomous criterion are computed from the cell entries of a fourfold table. The techniques applicable in this situation are either correlation-methods or difference-methods. The degree of relationship between success on the item and success on the criterion can be measured by a tetrachoric coefficient of correlation, a coefficient of colligation, or a phi coefficient.[1] The other methods depend upon the percent of the upper and the percent of the lower groups getting the item right. The simple difference between the two percentages is the easiest to compute but a chi-square comparison is preferable because it indicates the significance of a difference.[2]

The selection of test items with either a continuous score or a dichotomy requires a considerable amount of time for the computation. To reduce this computational time, extreme groups are used for item selection purposes. These short-cut methods economize on the time by sacrificing the quantitative nature of a continuous test score distribution. If the relationship of item score to test score is linear, so that the percentage of successes on the items increases as the total score increases, the differences on a single item between the upper and lower groups will

---

[1] P. E. Vernon, "Indices of Item Consistency and Validity," British Journal of Psychology, Statistical Section, I, 152-66, 1948.

[2] Ibid.

be sharpened by taking extreme groups. However, the increased sharpness of discrimination is somewhat offset by a loss of information which results from excluding some cases in the middle of the test score distribution.

The use of extreme groups necessitates balancing the sharpness of the discrimination and the stability of the indices. Kelley[1] has presented the mathematical proof that the upper and lower 27 percent of a sample are the optimum groups to use, provided the difference in the criterion scores among the members of each group is not utilized. The 27 percent maximizes the critical ratio based upon the difference between the means of the two groups. Each item and the criterion score are regarded as normally distributed and continuous variables. Kelley[2] also outlined a procedure for estimating a product-moment correlation coefficient between the item and the criterion score, excluding the item in question. Three techniques have been developed that are applicable where the two extreme groups each constitute 27 percent of the total group. A summary description of these methods follows.

(1) Biserial r (r_{bis}) approximation by Flanagan's method[3]

A table containing the values of the correlation coefficients in a normal bivariate surface corresponding to various combinations of propor-

---

[1] T. L. Kelley, "The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology, XXX, pp. 17-24, 1939.

[2] Ibid.

[3] J. C. Flanagan, "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from the Data at the Tails of the Distribution," Journal of Educational Psychology, XXX, pp. 674-80, 1939.

tions in the upper and lower 27 percent of the group was constructed. A normal bivariate surface assumes a normal distribution underlying both the dichotomous item response and the criterion variable; it also assumes rectilinearity of regression. To use this technique, it is necessary to determine the number of the high group that answered the item correctly and express this as the proportion of the high group; a similar number is obtained for the low group. These two proportions are looked up in the table and the approximate correlation coefficient is found.

(2) z method[1]

Since equal increments in r_{bis} do not represent equal increments in discriminating power, Davis transformed the r's into z's and converted z to a scale with a mean of fifty and a standard deviation of twenty-one. The z values may be added, subtracted, or averaged. A chart is provided from which one may read off the difficulty (expressed in sigma units) and the discrimination indices corresponding to various values of the upper and lower proportions of success. To use this chart the criterion scores and the percent knowing the correct answer to an item are corrected for chance.

(3) Probable error of percent difference

Votaw[2] and Arnold[3] gave formulae and nomographs for reading off the probable error of the percent difference in the upper and lower 27 percents.

---

[1]F. B. Davis, Item-Analysis Data: Their Computation, Interpretation and Use in Test Construction, (Harvard Education Papers, No. 2.), Cambridge: Graduate School of Education, Harvard University, 1946.

[2]D. F. Votaw, "Graphical Determination of Probable Error in Validation of Test Items," Journal of Educational Psychology, XXVI, 682-86, 1935.

[3]J. N. Arnold, "Nomogram for Determining Validity of Test Items," Journal of Educational Psychology, XXVI, 151-53, 1935.

The difference-methods subsumed under the rubric dichotomous groups[1] are applicable using the extreme 27 percents of the criterion group. The level of significance for the difference-method indices is dependent upon the number of cases in the groups; this makes it impossible to compare, by a statistical test, indices based on extreme groups with those based on dichotomous groups.

Other techniques of item validation have been proposed. One of these, the double tetrachoric index[2], is computed by dividing the criterion into three groups on the basis of the thirty-third and the sixty-sixth percentiles and averaging the tetrachoric correlations obtained by the two splits.

A simplified formula for the product-moment correlation coefficient of a dichotomous variable with a multiple-categoried variable, when the criterion is coded 2, 1, 0, -1, -2 to yield a rectangular distribution, is

$$r_{xy} = \frac{2a + b - d - 2e}{\sqrt{2k(n-k)}}$$

where n = number of persons, k = number of persons selecting the correct response, and a, b, c, d, and e respectively denote the frequency of correct choices by the five coded groups.[3]

Although many item selection techniques have been presented that are based upon the relationship between the item and the criterion, when one considers the computational time and the stability of the discrimination indices, Flanagan's method appears to be the most satisfactory one to use.

---

[1] P. 4.

[2] Vernon, loc. cit.

[3] D. C. Adkins and H. A. Toops, "Simplified Formulas for Item Selection and Construction," Psychometrika, II, 165-171, 1937.

Item Analysis to Maximize Validity

In most cases it is probable that the selection of items to increase reliability will increase validity; however, it has been demonstrated that it is possible to increase reliability and decrease validity or increase validity while decreasing reliability.[1] Thus it is found that mere selection of items correlating highest with an external criterion does not necessarily produce the most valid test. The ideal test is one composed of items which correlate highly with an external criterion and poorly with one another. Theoretically, if suitable external criterion scores were available, correlation coefficients between each item and the criterion could be obtained; the intercorrelations between test items could also be obtained. A multiple regression weight could be computed for each item and those items having regression weights significantly different from zero at a specified level of confidence could be selected for the final form of the test. Since the computations necessary to determine which combination of items would yield the largest multiple correlation coefficient is laborious, many approximation methods to this multiple regression problem have been suggested. The method of successive residuals[2] and the L-method[3] depend on building up successive composites of the most valid items; they require fewer item intercorrelations than the multiple

---

[1]H. E. Brogden, "Variation in Test Validity with Variations in the Distribution of Item Difficulties, Number of Items, and Degree of Their Intercorrelations," Psychometrika, XI, 197-214, 1946.

[2]A. P. Horst, "Item Analysis by the Method of Successive Residuals," Journal of Experimental Education, II, 254-63, 1934.

[3]H. A. Toops, "The L-Method," Psychometrika, VI, 249-66, 1941.

regression solution, but the task is a lengthy one. A presentation of these two techniques is given by the authors.

Richardson and Adkins[1] presented a simple approximation to the multiple correlation procedure that compared favorably with the L-method.[2] This formula is

$$r_{xy} = \frac{r_{yu} - r_{xu}\, r_{xy}}{\left(r_{xy} - r_{yu}\, r_{xu}\right)\sigma_u}$$

where y = criterion variable,

x = test variable, and

u = any test item.[3]

Flanagan[4] adapted the method of solving for regression coefficients by means of successive approximations to provide an item selection method.

A second technique proposed by Horst[5], the maximizing function, is dependent on the ratio of the validity of the item to the consistency of the item. Items are selected that correlate highly with the external criterion and poorly with the test score.

---

[1]M. W. Richardson, and D. C. Adkins, "A Rapid Method of Selecting Test Items," Journal of Educational Psychology, XXIX, 547-52, 1938.

[2]H. A. Toops, loc. cit.

[3]M. W. Richardson and Adkins, op. cit. p. 549.

[4]J. C. Flanagan, "A Short Method for Selecting the Best Combination of Test-Items for a Particular Purpose," Psychology Bulletin, XXXIII, 603-4, 1936. (Seen in abstract only).

[5]A. P. Horst, "Item Selection by Means of a Maximizing Function," Psychometrika, I, 229-44, 1936.

Use of Item Analysis Data

In the construction of a test, the distribution of test scores may be predetermined, within limits, by the proper selection of items with certain difficulty indices. The following formulae show that the sample mean and variance are functions of the item difficulty indices and the interaction between items.

$$\bar{X} = \sum_i p_i$$

$$\sigma^2 = \sum p_i q_i + 2 \sum_{\substack{i,j \\ i<j}} (p_{ij} - p_i p_j),$$

where p_i = the proportion of the individuals in the sample passing item i,

$q_i = 1 - p_i$, and

p_{ij} = the proportion of the individuals in the sample passing both items i and j.

The symmetry or asymmetry of the distribution of test scores, the skewness, kurtosis, or modality are also functions of the item difficulty indices and the item interactions.

Since the score distribution properties are dependent upon the item indices, items should be selected which will yield a score distribution best serving the purpose for which the test is to be used. No one frequency distribution exists which would be ideal for all testing situations. In objective testing, extensive use has been made of the normal test score distributions because: (1) the ability being measured was assumed to be normally distributed, and (2) the statistical methods applied in the theory of measurement are based on normal probability theory. The first reason is meaningless unless the ability is given an operational definition; normality is usually assumed on a philosophical basis. The second

reason is not relevant since the test was constructed for a specific purpose other than the application of statistical methods to the obtained data.

There are certain testing situations where the primary purpose is to classify individuals into two or more groups; no attempt is made to identify differences between the individuals within a particular group. Some personality testing is undertaken to measure the presence or the absence of a trait, with no attempt being made to measure the intensity of the trait. A test, best serving the purpose of this situation, would have sufficient accuracy at the critical point of partition to insure that the classification of a student into one of the two groups was not a result of chance fluctuations. In the measurement of interest, a test is desired which would classify individuals into certain interest groups. The test would need accuracy of distinction between groups rather than within groups. In achievement testing for the purpose of assigning grades on the basis of a specified point scale, a test should identify students receiving one grade from students receiving other grades. Test items selected by the item analysis techniques now in use generally result in a test score distribution not significantly different from a normal distribution. Consequently, various mathematical transformations are applied to the raw score data to obtain the critical points of partition between groups.

Purpose of This Study

This study was undertaken to develop and test a new technique for the selection of those items most applicable in testing situations where it is

desired to place individuals into a number of mutually exclusive groups. A technique was desired which would select the test items yielding a maximum discrimination between the groups. To obtain this maximum difference between the mean scores of the groups, it is necessary to maximize the ratio of the between groups variance to the total variance.

The theoretical aspects of the problem consist of the presentation of the technique and an analysis of its effectiveness. To test the worth of this technique an analysis of the resulting test score distribution is necessary; it is also necessary to compare the resulting distribution with a normal test score distribution to see which is more efficient in differentiating between two groups of individuals.

It is possible that a theoretical proof is true even though such a technique does not work in an actual situation. For this reason, some empirical data are necessary in order to determine the practical value of the technique. The primary need of any statistical technique is that it is stable and consistent when reapplied in another situation; to investigate the stability of the new technique it is necessary to use cross-validation procedures. Empirical data are necessary to investigate whether the new technique results in a score distribution with minimum error at the critical score points between two adjacent groups. Also, a comparison with other techniques seems desirable. Since it is not practical to compare the new procedure with all the present procedures, it will be compared with Flanagan's technique, which is based on item-criterion relationship only, and the method of maximizing function,[1] which is one of the better techniques based upon both the item-criterion and the item-item relationship.

---

[1] A. P. Horst, loc. cit.

CHAPTER II

THE THEORETICAL SOLUTION OF THE PROBLEM

The assumption underlying the placing of individuals within a specified grouping arrangement should be that there exists a real difference between the members of different groups and that the members of any one group are fairly homogeneous with respect to the trait being measured. Any test utilized for grouping purposes should yield an array of scores for a particular group that does not overlap the score distribution of any of the other groups. In the ideal case, with perfect items, individuals could be placed into m categories with m-1 items. Item one would be failed by group 1 and passed by groups 2, 3,...,m; item two would be failed by groups 1 and 2 and passed by groups 3, 4,...,m; item m-2 would be failed by groups 1, 2, 3,...,m-2 and passed by groups m-1 and m; the last item m-1 would be failed by all groups except group m. These m-1 perfect items would yield a score distribution where all the individuals in group 1 received a score of zero, those in group 2 a score of one, those in group m-1 a score of m-2, and those in group m a score of m-1. It is not practical to utilize a single item because (1) a single item is subject to fluctuation in response from trial to trial, and (2) the correlation between an item and the criterion being predicted by the total test is so low that an item curve is comparatively flat and not representative of the total test discrimination. The many factors that operate to reduce the efficiency of a test result in a greater likelihood of error in predicting when a single item is used than when the total test is used. Since perfection is not likely attainable, a number of items

that function within chance limits of a perfect item might serve the same purpose as one perfect item. To place students into m categories we would need m-1 groups of items with each group of items approximating a perfect test item functioning about a particular critical point separating two adjacent groups. A technique is proposed for the selection of test items when the trait being measured is on a continuum and the groups are separated by a specified number of critical points along the continuum. The application of this technique in the situation where the test is measuring traits not on a continuum but rather on two or more continua would possibly result in a single item's functioning at a critical point on more than one of the continua. In this situation it is necessary to score the tests on the basis of sub-parts identifiable with one of the continua; this could result in a single item's being included in the scoring arrangement of more than one sub-test. It should be noted that if a single trait is being measured, an item would function at only one critical point; an individual's performance on the test would be indicated by a single score. In the case of two or more continua, an item may function at one critical point on one or more continua; an individual's performance must be indicated by more than one score.

Procedure for Item Selection

According to the above discussion, the procedure for the selection of test items where individuals are to be classified into m groups would be as follows:

1. Classify the individuals into the proper one of the m groups on the basis of either an external or internal criterion.
2. Select a sample of size n_i from group i. The calculations will be simplified if all n_i are equal.

3. For each test item determine the number of each group that answered the item correctly.
4. On an a priori basis obtain the theoretical frequency of group i. This is obtained by assuming a perfect test item so all n_i people of a group would pass the item if that group were above the critical point at which the item was discriminating, and all n_i people of a group below the critical point would fail the item.
5. The observed frequency is obtained by examining the number of each group that answered the item correctly. If the group is above the critical point the observed frequency is equal to the number of the group that answered the item correctly. The number of the group that failed the item we will denote by e_i. The observed frequency of successful predictions by the item for the group i is then equal to n_i minus e_i. If the group is below the critical point we would predict that all the members of the group would fail the item. The number of errors is thus equal to the number of the group that answered the item correctly. If we also designate this number by e_i, the observed frequency of successful predictions is also equal to n_i minus e_i.
6. Using the above theoretical and observed frequencies we use chi-square to test whether the observed frequencies deviate significantly from the theoretical frequencies.
7. By specifying the chi-square limits of acceptance and rejectance, we can identify the items that are acceptable at the various critical points.
8. If we restrict ourselves to chance deviations from the theoretical frequencies, the value of e_i is equal to qn_i, where q is the probability of getting an item right on the basis of chance alone. For a test item of a alternatives q is equal to 1/a. Using these limits we have an acceptance point for chi-square equal to $\sum_i q^2 n_i$.

The chi-square value is obtained by the formula

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where fo observed frequency,
and fe expected frequency.

An example of the procedure in the case of five groups is presented for clarification purposes. Let I denote the highest group and V the lowest and have all n_i equal to 10.

	Groups (from high to low)				
	I	II	III	IV	V
Number in each group	10	10	10	10	10
Number in each group answering the items correctly	10	8	4	4	1

This item tends to separate groups I and II from groups III, IV, and V. For a perfect test item operating at this critical point we would expect all the individuals of groups I and II to pass and the other individuals to fail. The chi-square value is calculated as follows:

	Groups (from high to low) I	II	III	IV	V
fe	10	10	10	10	10
fo	10	8	6	6	9
fo-fe	0	2	4	4	1
$(fo-fe)^2$	0	4	16	16	1
$\frac{(fo-fe)^2}{fe}$	.00	.40	1.60	1.60	.10

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 3.70.$$

It is unlikely that we can find a sufficient number of items which will satisfy the chance deviation limits; but we may then use some other chi-square value, based on predetermined levels of significance, for the acceptance point. A satisfactory item has a chi-square value less than the acceptance value.

Theoretical Analysis of the Score Distribution in the Case of Two Groups

Let us denote the higher group by 1 and the lower group by 2. Consider a test consisting of k items with a alternatives and assume that the credit given is either 1 or 0 depending on whether the response is correct or incorrect. Let us further assume that for each item the probability of success is $\geq (a-1)/a$ for group 1 and $\leq 1/a$ for group 2. If we denote the probability of success for group 1 on item i by p_i, we consider

the case of k trials with different probabilities of success p_i, where $i = 1, 2, \ldots, k$.

Aitken[1] has shown that this type of distribution has

(1) $\sigma_1^2 = \sum p_i q_i$ where $q_i = 1 - p_i$.

Let p represent the mean probability of success for group 1, $p = \frac{1}{k}\sum_i p_i$; the variance of the probability in the k trials is

(2) $\sigma_p^2 = \sum_i (p_i - p)^2 / k$.

We have

(3) $\sum_i p_i q_i = kp - kp^2 - \sum_i (p_i - p)^2$.

Substituting the value for the last term on the right from equation (2), we have

(4) $\sum_i p_i q_i = kp - kp^2 - k\sigma_p^2$.

Simplifying equation (4), we obtain

(5) $\sum_i p_i q_i = kpq - k\sigma_p^2$.

Substituting from equation (5) in equation (1), we have

(6) $\sigma_1^2 = kpq - k\sigma_p^2$.

It is apparent that σ_1^2 will be a maximum for a mean probability, when $\sigma_p^2 = 0$. If we let $\sigma_p^2 = 0$, it follows that

(7) $\sum_i (p_i - p)^2 = 0$,

Hence $p_i = p$ for all i.

In a similar manner it may be shown that

(8) $\sigma_2^2 = kpq - k\sigma_p^2$, where p is the mean probability of group 2, and σ_2^2 is a maximum when $p_i = p$ for all i.

---

[1] A. C. Aitken, Statistical Mathematics, 2nd Edition, Interscience Publishers Inc., New York, N. Y., 1942, pp. 50-51.

In the theoretical analysis of the problem we will choose P_i so that the maximum variance is obtained, which means that $P_i = (a-1)/a$, $i = 1, 2, \cdots, K$, for group 1 and $P_i = 1/a$, $i = 1, 2, \cdots, K$, for group 2.

The distribution of scores for group 1 is characterized by

$$(9) \qquad \bar{X}_1 = \frac{K(a-1)}{a}, \quad S_1^2 = \frac{K(a-1)}{a^2}.$$

The mean and variance for group 2 is

$$(10) \qquad \bar{X}_2 = \frac{K}{a}, \quad S_2^2 = \frac{K(a-1)}{a^2}.$$

It is now necessary to determine whether the observed score distributions, identified by equations 9 and 10, classify the individuals into one of two groups with a small probability of error. The exact amount of error could be determined if the true scores for all individuals were known. Since true scores are unattainable, reasonable limits for the difference between the true scores of the individuals must be expressed in terms of the observed scores. To derive the relationship between true score differences and observed score differences, it is necessary to make some assumptions regarding the relationship between observed scores and true scores. The relationship between the observed scores, true scores, and error scores is assumed to be

$$(11) \qquad x_i = t_i + e_i$$

where x_i = observed deviation score of individual i,

t_i = true deviation score of individual i, and

e_i = error deviation component of individual i.

All errors are assumed to be random errors and are such that

$$(12) \qquad \bar{e} = 0$$

$$r_{te} = 0$$

$$r_{e_i e_j} = 0.$$

In the derivation of the relationship between true score and observed score differences, the summation index is omitted for ease of presentation. The summation is over i unless otherwise stated.

The relationship between the variance of the true, the error, and the observed scores is determined. From equation (11), we have

(13) $$x = t + e .$$

Squaring and summing gives

(14) $$\Sigma x^2 = \Sigma t^2 + \Sigma e^2 + 2\Sigma et .$$

Dividing both sides by N, we have

(15) $$S_x^2 = S_t^2 + S_e^2 + 2 r_{te} S_e S_t .$$

From (12), we see that the last term of equation (15) is zero, and we have

(16) $$S_x^2 = S_t^2 + S_e^2 .$$

We may solve for the variance of true scores in terms of the reliability of the test and the variance of the observed scores. The correlation between two parallel tests is defined, from elementary statistics, as

(17) $$r_{12} = \frac{\Sigma x_1 x_2}{N s_1 s_2}$$

where x_1 and x_2 are the scores of an individual on tests 1 and 2, and s_1 and s_2 are the variance on tests 1 and 2.

From (13) we may express the numerator on the right side of (17) as follows,

(18) $$\Sigma x_1 x_2 = \Sigma (t_1 + e_1)(t_2 + e_2) .$$

Expanding equation (18) gives

(19) $$\Sigma x_1 x_2 = \Sigma t_1 t_2 + \Sigma e_1 t_2 + \Sigma e_2 t_1 + \Sigma e_1 e_2 .$$

From the definitions of (12) we see that the last three terms of equation (19) are each zero. Since we have parallel tests, the true score on 1 and the true score on 2 are equal. Therefore equation (19) becomes

(20) $$\Sigma x_1 x_2 = \Sigma t_1^2 .$$

We may divide both sides of equation (20) by N, and from the definition of a variance we see that

$$\frac{\sum x_1 x_2}{N} = s_t^2. \tag{21}$$

Substituting equation (21) in equation (17) gives

$$r_{12} = \frac{s_t^2}{s_1 s_2}. \tag{22}$$

Since tests 1 and 2 are parallel, $s_1 = s_2$ and we see that

$$s_t^2 = r_{12}\, s_x^2, \tag{23}$$

where $s_x = s_1 = s_2$.

Since the reliability of a test, r_{xx}, is defined as the correlation between two parallel tests, we have

$$s_t^2 = r_{xx}\, s_x^2. \tag{24}$$

Next we may solve for the error variance by substituting equation (24) in equation (16), obtaining

$$s_x^2 = s_e^2 + r_{xx}\, s_x^2. \tag{25}$$

Solving equation (25) for s_e^2 gives

$$s_e^2 = s_x^2 (1 - r_{xx}). \tag{26}$$

It is necessary to determine the standard error of the difference between two scores, $x_i - x_j$. To write the formula for this error, we use equation (11) and write

$$x_i - x_j = t_i - t_j + (e_i - e_j). \tag{27}$$

The term in the parentheses indicates the error. The variation of the observed difference from the true difference is denoted by

$$\sum (e_i - e_j)^2 = \sum e_i^2 + \sum e_j^2 + 2 \sum e_i e_j. \tag{28}$$

From (12), we see that the last term of equation (28) is zero. Substituting equation (26) in equation (28), we have

$$(29) \qquad \Sigma (e_i - e_j)^2 = 2N S_x^2 (1 - r_{xx}).$$

Dividing by N and taking the square root, we have

$$(30) \qquad S_{(e_i - e_j)} = S_x \sqrt{2} \sqrt{1 - r_{xx}} .$$

It should be noted that in the development of the equation for the standard error of a difference, no assumptions were made regarding the distribution of errors. However, in order to utilize this error to obtain reasonable limits for the value of the difference between true scores, some assumption regarding the frequency distribution of errors is necessary. Let us make the usual assumption that the distribution of errors is normal.

For two individuals with a given score difference $x_i - x_j$, reasonable limits for the difference of true scores, $t_i - t_j$, may be taken as

$$(31) \quad x_i - x_j + 3\sqrt{2}\, S_x \sqrt{1 - r_{xx}} > t_i - t_j > x_i - x_j - 3\sqrt{2}\, S_x \sqrt{1 - r_{xx}} .$$

If the above limits include zero, there is no significant difference between t_i and t_j since $t_i - t_j$ may be zero. Since true differences were assumed to exist between the individuals of the different groups, that is $t_i - t_j > 0$, it follows that both of the limits of the above inequality must be positive and

$$(32) \qquad t_i - t_j > x_i - x_j - 3 S_x \sqrt{2} \sqrt{1 - r_{xx}} \geq 0.$$

Hence it is necessary that

$$(33) \qquad x_i - x_j - 3 S_x \sqrt{2} \sqrt{1 - r_{xx}} \geq 0$$

to be certain that true score differences exist between the individuals of group 1 and those of group 2.

If our selected test items were perfect for group 1 and chance operated for group 2, the groups would have the following means and variances:

(34)

x	Group 1 k	Group 2 k/a
s_x^2	0	$(a-1)k/a^2$.

It is quite apparent that no individual was placed in the wrong group on the basis of chance errors.

The k test items were selected so that the two groups have the means and variances given in (9) and (10). Since we have assumed that a real difference exists between the individuals of group 1 and the ones in group 2, we are saying the true scores of the individual at the low extreme of the group 1 distribution is greater than the true score of the highest scoring individual in the other group. If we let i represent the individual from the high group and j the one from the low group, we can express our assumption as $t_i - t_j > 0$. For our distributions we may be certain that

(35) $$X_i \geq \frac{(a-1)k}{a} - \frac{3}{a}\sqrt{k(a-1)}\,, \qquad \text{and}$$

$$X_j \leq \frac{k}{a} + \frac{3}{a}\sqrt{k(a-1)}\,.$$

Taking the maximum value for x_j and the minimum value for x_i and substituting these values in (33), we have

(36) $$\frac{(a-1)k}{a} - \frac{3}{a}\sqrt{k(a-1)} - \frac{k}{a} - \frac{3}{a}\sqrt{k(a-1)} - 3S_x\sqrt{2}\sqrt{1-r_{xx}} \geq 0.$$

Multiplying both sides of the inequality by a and combining like terms, we have

(37) $$(a-2)k \geq \sqrt{k(a-1)}\left[6 + 3\sqrt{2}\sqrt{1-r_{xx}}\right].$$

Squaring both sides and simplifying, we have

$$(38)\quad k \geq \frac{(a-1)}{(a-2)^2}\left[6 + 3\sqrt{2}\,\sqrt{1-r_{xx}}\right]^2, \quad a > 2.$$

Considering the quantity $\sqrt{1-r_{xx}}$ as one variable and k as the other variable, the inequality (38) yields boundary values that represent a family of parabolas dependent upon the parameter a. The value of k must always lie in the positive quadrant because positive square roots are taken. Since $0 \leq r_{xx} \leq 1.00$, we are concerned only with the half parabolas. The value of k which will meet the equality of (38) for a fixed a lies on the curve defined by (38); all k greater than this value of k will satisfy (38). Figure 1 indicates the number of test items needed for various combinations of test reliabilities and the number of alternatives for each test item.

An example is given to illustrate the method of reading Figure 1. A four-choice item test with a reliability of .20 would require 72 items to efficiently separate the individuals into two groups; a three-choice item test would need a reliability of 1.00 to accomplish the same end. From Figure 1 it appears that the number of alternatives each item has is an important factor in the number of items required to perform the discrimination between the two groups; however, the test will usually have a reliability greater than .50, so it is apparent that the number of alternatives is not of great importance when the number of alternatives is five or more.

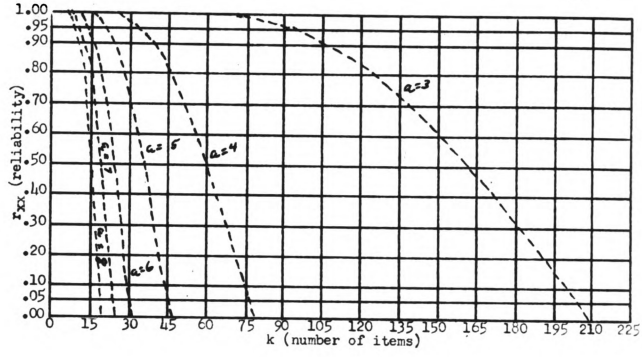


FIGURE 1. The Number of Items Required to Make a Discrimination Between Two Groups as a Function of the Test Reliability and the Number of Alternatives for Each Item.

It follows that for any test of reliability $r_{xx'}$ we may be reasonably certain, probability of 9,987.5 in 10,000, that $t_i - t_j > 0$ by selecting k large enough so that (38) is satisfied. It is apparent that if the items were more homogeneous for group 1, the variance of the group would become smaller and the mean would approach k.

Comparison of the Bimodal Distribution and the Normal Distribution

When the probability of success of the group is a constant for each item, the scores for the group will form a Bernoulli distribution. The theoretical relative frequencies for the dichotomous situation are given by the terms of $(p+q)^k$. This type of distribution is characterized by

the following functions[1]

(40)
$$\bar{X} = kP$$
$$\sigma^2 = kPq$$
$$\alpha_3 = (q-P)/\sqrt{kPq}$$
$$\alpha_4 = \frac{1}{kPq} - \frac{6}{\sqrt{kPq}} + 3.$$

The skewness is positive for $p < 1/2$, negative for $p > 1/2$, and zero for $p = 1/2$. As k approaches infinity, α_3 tends to zero and α_4 tends to 3.

In the comparison let the range of the distributions be from 0 to h; and let the bimodal distribution be such that $\bar{x}_1 - 3\sqrt{k(a-1)}/a = \bar{x}_2 + 3\sqrt{k(a-1)}/a$, which means that the score distributions for the two groups intersect at a point. This point is the weighted average of their respective means; when the size of the two groups is equal the point of intersection is equidistant between the two means. For the normal distribution we have $x = h/2$, and $s_x = h/6$.

Let us assume that for the bimodal distribution we have a standard error of measurement equal to the standard deviation of the group, or assume the reliability of the test is zero for the group. Let us also assume the entire distribution of each of the two groups in the bimodal case lies in the interval $\bar{x}_i \pm 3s_i, (i = 1,2)$. For the normal distribution nearly all the cases lie in the interval $\bar{x} \pm 3s_x$. If we define a critical region about the critical point of partition between the two groups, we would have a certain percentage of the cases of each bimodal group within this band. For our assumed error of measurement equal to the standard

---

[1] Aitken, op. cit., pp. 49-50.

deviation of one of the groups, the area under the curve for group 1 that is within this critical region is equal to the area under the normal probability curve between the t values of 2 sigma and 3 sigma. This area is equal to .023 of the total area, and it follows that 2.3% of the individuals in group 1 lie within this critical region. The percentage of the individuals in group 2 that are within this region is also equal to 2.3% of the group. For the combined total 2.3% of all of the individuals lie within this error band. For the normal distribution to have as small a percentage of the cases within a critical region, it is necessary that

(41) $$\frac{1}{\sqrt{2\pi}}\int_{c-\sigma_t}^{c+\sigma_t} e^{-\frac{t^2}{2}}dt = .023,$$ where c is the critical point depending upon the proportions in each of the two parts of the normal curve.

It is apparent from the Table of the Normal Curve and the standard error of measurement of a standard score[1] that the reliability of the test, necessary to obtain this accuracy, is dependent upon the point of partition between the two groups. If the point of partition is in the tail of the normal distribution at a t value of 2.00, this accuracy is attainable only if the reliability of the test is greater than .91. As the point of partition approaches the mean of the normal distribution, the reliability of the test must increase to maintain the same degree of accuracy; in the limiting case with a t value of .00, the reliability of the test must be at least equal to .9991.

If we assume that the two distributions have equal errors of meas-

---

[1] Standard error of measurement of a standard score $\sigma_t = \sqrt{1-r_{xx}}$.

urement, we have the following areas under the curve in the critical region

$$(42)\quad \text{Bimodal case } \frac{1}{\sigma_1\sqrt{2\pi}}\int_{-3+S\cdot E.}^{-3.00} e^{-\frac{(x_1-\bar{x}_1)^2}{2\sigma_1^2}}\,dx_1 + \frac{1}{\sigma_2\sqrt{2\pi}}\int_{3-S\cdot E.}^{3.00} e^{-\frac{(x_2-\bar{x}_2)^2}{2\sigma_2^2}}\,dx_2$$

$$\text{Normal case } \frac{1}{\sqrt{2\pi}}\int_{c-S\cdot E}^{c+S\cdot E} e^{-\frac{t^2}{2}}\,dt$$

where c is the point of partition between the two groups.

From the properties of the normal curve, it is immediately seen that

$$(43)\quad \frac{1}{\sigma_1\sqrt{2\pi}}\int_{-3+S\cdot E}^{-3.00} e^{-\frac{(x_1-\bar{x}_1)^2}{2\sigma_1^2}}\,dx_1 + \frac{1}{\sigma_2\sqrt{2\pi}}\int_{3-S\cdot E}^{3.00} e^{-\frac{(x_2-\bar{x}_2)^2}{2\sigma_2^2}}\,dx_2 < \frac{1}{\sqrt{2\pi}}\int_{c-S\cdot E.}^{c+S\cdot E} e^{-\frac{t^2}{2}}\,dt.$$

for all values of c.

The preceding theoretical development indicates a bimodal test score distribution is superior to a normal test score distribution in the classification of individuals into two groups. To minimize the occurance of chance errors in this type of testing situation, a bimodal test score distribution, with a critical point of partition at the minimal ordinate between the modes of the two groups, should be used. A similar proof will show that an m-modal distribution will best serve the purpose of classifying the individuals into m groups. It is apparent that an increase in the number of items needed to adequately perform the job of classification is necessary when more than two groups are used.

CHAPTER III

DATA RELATED TO THE PROBLEM

To test the adequacy with which the adjacent-group item selection technique functions, two studies directed at the empirical verification of the technique were undertaken. The first investigated the stability of items selected by this technique under cross validation. The second study compared the adjacent-group technique with a technique based on item-criterion relationship and a technique based on both item-criterion and inter-item relationships. Comprehensive achievement tests given in courses of the Basic College at Michigan State College were used in these empirical studies.

Stability of the Selected Items under Cross-Validation

Any satisfactory item selection technique must yield items that will be valid when applied to similar populations. This study was undertaken to investigate whether the adjacent-group technique would yield a bimodal distribution for two groups that was consistent when used with another independent sample.

The Test. The adjacent-group item selection technique was applied to an achievement examination. The examination consisted of three hundred items; the majority of the items were of the five response, multiple choice variety, of which only one response was considered correct. The questions were scored one for a correct response and zero for an incorrect response or omitted item. The time limit was sufficient to allow everyone adequate time to attempt all items.

Selection of the Items. The 1,782 individuals taking the test were divided into five groups according to a letter grade of A, B, C, D, or F on the basis of the score on the 300-item examination. A sample of 50 individuals was obtained by selecting ten individuals from each of the five strata A, B, C, D, and F. For the five groups, items were analyzed to determine the number in each group that answered each item correctly. From this data chi-square values were computed for each of the 300 items. On the basis of chi-square acceptance values, three sub-tests were formed. The first consisted of 59 items emphasizing discrimination between the B and C groups; the second consisted of the 12 best items for discriminating between the two lower categories; and the third consisted of these 12 plus the next best 16 items discriminating between the D and F groups.

Consistency Data. For the cross-validation study a stratified random sample of 199 papers was selected from the total group of 1,782 papers. For the first sub-test a score of 190 on the original 300-item test was the minimum of the B letter grade group. Since the original test had an error of measurement of 6.84, all the scores of the other groups should be below 162 before we could be certain real differences existed between the two groups.[1] From the sample of 199 individuals, we obtained 62 people with a score equal to or greater than 190 and 57 people with a score of 161 or less. These 119 papers were rescored on the basis of the 59 selected items. The mean and range of scores for each group are given in Table I.

---

[1] P. 22, equation (33).

TABLE I. MEANS, RANGES, AND FREQUENCIES FOR THE TWO GROUPS ON THE 59-ITEM TEST

Group	N	Range	$\bar{x}$
Original score greater than or equal to 190	62	36-52	45.79
Original score less than or equal to 161	57	19-35	29.05

The data presented in Table I indicate the selected test items resulted in a non-overlapping bimodal distribution.

The second part of the consistency study was carried out with the 199 individuals placed into a pass-fail dichotomy. The pass group consisted of 192 individuals receiving a letter grade of A, B, C, or D, and the seven students receiving a letter grade of F made up the failing group. These papers were scored on the basis of the 12- and 28-item sub-tests. The frequency distribution for the pass and fail groups for the selection sample and the cross-validation sample on the 12-item test are given in Table II.

TABLE II. FREQUENCY DISTRIBUTION OF EACH SAMPLE ON THE 12-ITEM TEST

Raw Score	Selection Sample		Cross-validation Sample	
	pass	fail	pass	fail
12	14		42	
11	10		60	
10	8		42	
9	8		24	1
8			9	
7			10	1
6		1	3	4
5		3	1	
4			1	1
3		2		
2		3		
1		1		

For the selection group a marked non-overlapping bimodal distribution was obtained. For the cross-validation sample the distributions for the pass and fail group overlapped. An investigation was made of the original test scores of the passing group individuals with a sub-test score of six or less and the failing individuals with a sub-test score of seven or more. The original scores for the incorrectly classified individuals are given in Table III.

TABLE III. ORIGINAL TEST SCORES FOR THE INDIVIDUALS INCORRECTLY PLACED ON THE 12-ITEM TEST

Group	Original Score
Passing original test and failing the 12-item sub-test	150 145 136 132 130
Failing original test and passing the 12-item sub-test	123 120

Reasonable limits for observed score differences, in order to be certain that true score differences exist, may be obtained from the original test.[1] This implies that to have a criterion score different from a criterion score of 128 an individual must have a total score of 157 or greater.[2] Of the individuals incorrectly classified by the 12-item test no differences existed on the original test of a magnitude sufficient to be reasonably certain that true score differences did exist.

---

[1] P. 22, equation (33).

[2] Probability of 9,987.5 in 10,000.

For the 28-item test a similar analysis was undertaken and 17 individuals were within the fringe area. The original criterion scores for these individuals are given in Table IV.

TABLE IV. ORIGINAL TEST SCORES FOR THE INDIVIDUALS INCORRECTLY PLACED ON THE 28-ITEM TEST

Group	Original Score	
Passing original and failing 28-item sub-test	169	144
	165	133
	152	132
	150	130
	147	130
	146	129
	145	
Failing original and passing 28-item test	128	125
	126	120

From Table IV we see that two individuals in the high group were incorrectly classified when we restrict ourselves to those individuals with true score differences. On the basis of the obtained data it appears that the adjacent-group technique does result in bimodal test score distributions, and it is consistent provided the items do not deviate too greatly from perfect items. The 12-item test and the 59-item test consisted of those items having a chi-square value less than 4.00; the 28-item test consisted of some with chi-square values greater than 4.00. It is apparent from the data that selecting additional items with large chi-square values causes a decrease in the accuracy of the test and a corresponding increase in the number of errors.

Although the theory covered only the case of two groups, this study was undertaken to investigate whether a multimodal test score distribution was attainable by the group technique; if it were attainable, the

multimodal test score distribution would be investigated with respect to the variation of its standard error of measurement as the magnitude of the test score changed. It seemed desirable to obtain items which groups above the critical point would pass with a probability as near to 1.00 as possible and the groups below the critical point would pass with a probability of .20 or less. A pool of 900 items was secured by combining three 300-item achievement tests. Since only 110 individuals were available who had taken all three tests, it was necessary to select test items without using these cases in the item selection process.

Individuals taking each test were classified into five grade groups, A, B, C, D, or F on the basis of total test score. Four critical points determined the groups. For each of the three 300-item tests, a sample of ten individuals was selected from each of the five groups, and a chi-square value was computed for each of the items. From these 900 chi-square values, a test of 60 items was selected on the basis of the adjacent-group technique; and a second set of items was selected on the same basis to yield a statistically parallel test.

The 110 individuals were divided into five groups on the basis of their average for the three tests. Table V presents the means, variances, and frequencies for each group on the two parallel tests and a total test score for individuals obtained by summing the part scores for an individual.

TABLE V. MEANS, VARIANCES, AND FREQUENCIES FOR THE FIVE GROUPS ON THREE TESTS

Groups	N	Parallel Test 1		Parallel Test 2		Total Test	
		$\bar{x}$	s^2	$\bar{x}$	s^2	$\bar{x}$	s^2
1	6	54.00	4.67	53.00	10.33	107.00	23.67
2	19	44.16	8.88	43.16	5.38	87.32	14.43
3	32	33.38	7.29	34.06	5.96	67.44	14.72
4	35	24.63	6.55	24.23	6.92	48.86	13.56
5	18	14.94	5.29	15.09	6.25	30.33	17.54

To compute the standard error of measurement for a test, on which the total score is the sum of the scores on two parallel tests, one may take the sum of the squares of the differences between corresponding individual scores, divide by the number of individuals, and extract the square root. In symbols,

$$(1) \qquad \sigma_{S\cdot E} = \sqrt{\sum \frac{(X_{i1} - X_{i2})^2}{n}}$$

where X_{i1} = the observed score of individual i on test 1,
X_{i2} = the observed score of individual i on test 2, and
n = the number of individuals.

An indication of the magnitude of the standard error of measurement at several points on the test score scale is provided by the means of the squared differences for individual scores. The means of the squares of differences between parallel test scores were calculated at each score point along the total test score scale. To secure a somewhat more stable value, the means of the squared differences were computed for groups of five score points along the total score axis. The values of the means of the squared differences obtained from the groups of five score points is not too stable since in the eighteen groups the greatest frequency for any group was seventeen and six of the groups consisted of only two or three individuals. The frequency distribution of the total score and the means of the squared differences for each score point and each group of five score points are presented in Figure 2. The empirical curve best fitting the data on the mean of the squared differences is also shown.

The four critical points of partition separating the five groups are located at the total raw scores of 40, 60, 80, and 100; and the fre-

quency at these points is zero. An examination of Figure 2 shows the curve of the mean squared difference of the multimodal score distribution is also multimodal in nature and the error of measurement is a minimum at the critical score points. This minimum at these points results in a smaller percentage of the cases being within the critical region about the four critical points. It is obvious that a distribution which yields a minimum standard error of measurement at critical points is the type of distribution that best fits the purpose of assigning individuals into groups separated by critical points and best meets the assumption that there is a real difference between the individuals of two adjacent groups.

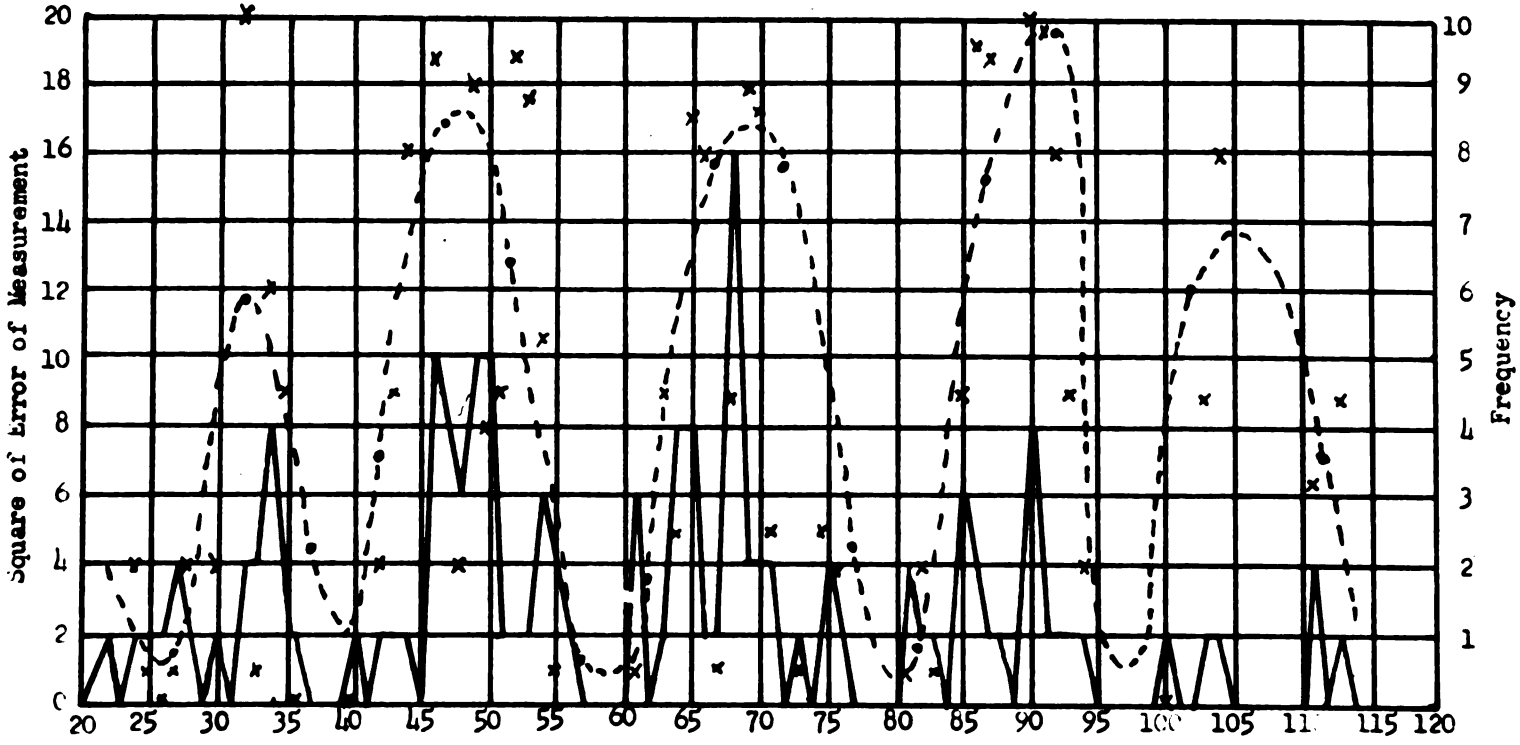
Comparison of the Adjacent-Group Technique and Two Other Techniques

This empirical study was undertaken to compare the adjacent-group technique of item selection and a technique based on item-criterion relationship with respect to the reliability and the validity of the subtests selected by each method.

The Test. Items were selected from an achievement examination in an area. The test consisted of 300 items; the majority of the items were of the five response, multiple choice variety of which only one response was considered correct. The questions were scored one for a correct response and zero for an incorrect one; omitted items were counted as wrong responses. The time limit of the test was sufficient to allow everyone adequate time to attempt all items.

Selection of the Items. A stratified random sample of 185 papers was selected for use with the upper-lower 27% technique.[1] The highest and low-

[1]Flanagan, loc. cit.



Legend

X mean of the squared differences for an individual score point

● mean of the squared difference for groups of five score points

frequency polygon of total score

- - - - - empirical curve best fitting the data on the mean of the squared differences

FIGURE 2. The Frequency Distribution of the Total Score and the Magnitude of the Means of the Squared Differences Throughout the Total Score Range

est 27% of the 185 papers were selected as criterion groups (50 papers in each group). Each of the 300 items was analyzed by the IBM Graphic Item Counter to determine the number of the high and the low 27% selecting the correct response. Two indices were computed for each item: (1) a difficulty index that was the percentage of the combined criterion groups missing the item, and (2) an index of discrimination computed from a chart yielding an approximate biserial correlation coefficient estimated from obtained proportions of the upper and lower 27%.

For the adjacent-group technique the 1,782 individuals were placed into five groups on the basis of their obtained grade and from each of these five groups a sample of ten individuals was selected. The 300 items were analyzed to determine the number in each group answering the item correctly; from these data a chi-square value was computed for each item.

Reliability and Validity Indices. After the item indices were obtained, four tests were made as a result of the analyses: (1) the best 31 items selected by the upper-lower method, (2) the best 31 items as indicated by the adjacent-group technique, (3) the best 59 items selected by the upper-lower method, and (4) the best 59 items selected by the adjacent-group technique.

A new stratified random sample of 199 papers was selected and rescored on the basis of the four new tests. For each of the four subtests a reliability estimate was obtained by the following formula[1]

(2) $$r_{xx} = \left[Ks^2 - \bar{x}(K-\bar{x})\right] / (k-1)\, s^2,$$

where k = number of items,
s^2 = variance of the test scores,
$\bar{x}$ = mean of the test scores.

---

[1] H. E. Garrett, Statistics in Psychology and Education. Third Edition. New York: Longmans, Green and Company. 1947, p. 385.

TABLE VI. RELIABILITY ESTIMATES FOR THE FOUR TESTS

Number of Items in Test	Items Selected by	Reliability Estimate
31	upper-lower	.64
31	adjacent-group	.58
59	upper-lower	.77
59	adjacent-group	.78

It is quite apparent that the reliability estimates for the two 31-item tests are not significantly different; no significant difference exists between the estimates for the 59-item tests.

Validity estimates were obtained by correlating the scores on the sub-tests with the original test scores. The obtained correlations for the two 31-item tests are tested for the significance of the difference in Table VII.

TABLE VII. TEST OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN THE VALIDITY COEFFICIENT ESTIMATES OF THE TWO 31-ITEM TESTS

Selection Technique	r	z	1/(n-3)
adjacent-group	.853	1.271	.00510
upper-lower	.775	1.034	.00510
	difference	.237	sum .01020

$s_d = \sqrt{.01020} = .101$, $t = .237/.101 = 2.35$ d.f. $= \infty$ $P = .02$.

The obtained difference between the two r's in favor of the adjacent-group selection technique could have occurred by chance factors alone about two times in one hundred. A similar test for the validity coefficients of the 59-item tests is given in Table VIII.

TABLE VIII. TEST OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN THE VALIDITY COEFFICIENT ESTIMATES OF THE TWO 59-ITEM TESTS

Selection Technique	r	z	1/(n-3)
adjacent-group	.91	1.535	.00510
upper-lower	.88	1.344	.00510
	difference	.189	sum .01020

$s_d = .101$, $t = .189/.101 = 1.87$, d.f. $= \infty$, $P = .06$.

For the two 59-item tests, the obtained difference in favor of the adjacent-group technique could have been caused by chance factors alone in about six cases in one hundred.

This study indicates that the adjacent-group technique will probably yield higher validity coefficients than the upper-lower technique, although there may be no appreciable difference in the reliability coefficients.

The most satisfactory method of selecting test items from a theoretical approach, is to consider the selection process as being similar to a multiple regression problem. The solution of a multiple regression problem in the case of a large number of variables is prohibitive; in test-item selection the number of variables would generally be greater than one hundred. To approximate the regression solution numerous techniques have been proposed; these techniques have been mentioned before.[1] Since these methods are the most efficient for the selection of the better items, the adjacent-group technique presented in this study was compared with one of them, namely, Horst's maximizing function.[2] A sub-test was constructed from the best 30 items as identified by the adjacent-group technique; a second sub-test consisted of the best 30 items selected by the maximizing function.[3] Estimates of reliability and validity were obtained for each of the two sub-tests. These estimates were tested for significance of the difference.

The Test. The test items were selected from an achievement examination consisting of 150 items (Form I); the majority of the items were of the

---

[1] P. 9-10.

[2] Horst, loc. cit.

[3] Horst, ibid.

five response, multiple choice variety of which only one of the responses was considered correct. One score point was given for a correct response and no points for an incorrect one. Omitted items were considered wrong since adequate time was given to enable all individuals to finish the test. A second test (Form II) over the same subject matter field was used as an external criterion. The correlation between the external criterion and the test was .68. Two stratified random samples were selected; one sample of 100 papers was used to select the test items for a 30-item test using the maximizing function and a 30-item test using the group method; the second was reserved for the computation of the reliability and validity estimates.

Selection of the Items. The mathematical theory underlying the selection of test items by the method of maximizing function was developed in the article by Horst[1] and will not be discussed in this paper. The computational procedure followed to select the best items by this method is:

1. Reduce both raw criterion and raw test score to class interval values ranging from 0 to 9.
2. For each item obtain the sum of the criterion measures (ΣY) and the sum of the test measures (ΣS) for all individuals answering the item correctly.
3. Next calculate the mean of the criterion measures (M_y) and the mean of the total test score (M_s).
4. For each item multiply the mean criterion score and the mean test score by the number of the people answering the item correctly. Indicate these by fM_y and fM_s.
5. Subtract fM_y from ΣY and designate the result by u. Subtract fM_s from ΣS and designate the result by v. The u value is proportional to the product moment of the corresponding item with the criterion.
6. If the u value is negative, discard the item since it correlates negatively with the criterion. Items having positive u values and negative v values are selected for the test.

---

[1] Horst, loc. cit.

7. For all the items not previously accepted or rejected divide u by v.
8. Find the highest u value and the highest v value. Divide the u by the v to obtain a ratio. Divide this ratio by 2.
9. Select all items greater than the ratio.
10. Make a frequency distribution of all u/v ratios which lie between the ratio and the ratio divided by 2. Select as many of the remaining items as are necessary to obtain a test of the desired number of items.

For the adjacent-group technique, the 100 papers were divided into five groups on the basis of four critical points of the external criterion. For each group the number getting each item correct was obtained and a chi-square value was computed for each item.

Reliability and Validity. After the two analysis methods had been completed, two tests of 30 items were chosen. The one test was made up of the best 30 items selected by Horst's method; the other consisted of the best 30 items as shown by the adjacent-group method. For each test an estimate of the reliability was obtained using the second sample and is presented in Table IX.

TABLE IX. RELIABILITY ESTIMATES FOR THE TWO 30-ITEM TESTS

Method of Item Selection	Reliability
Horst	.47
Group	.52

These two estimates are obviously not significantly different. The validity estimates, based on the second sample, were obtained by correlating the scores on the sub-tests with the criterion scores (Form II) and are presented in Table X.

TABLE X. VALIDITY ESTIMATES FOR THE TWO 30-ITEM TESTS

Method of Item Selection	Validity
Horst	.54
Group	.58

These two correlation coefficients are also not significantly different.

The validity estimates of the original test (Form I) and the 30-item sub-test consisting of items selected from Form I by the adjacent-group method were also tested for the significance of the difference. The results are given in Table XI.

TABLE XI. TEST OF SIGNIFICANCE OF THE DIFFERENCE BETWEEN THE VALIDITY COEFFICIENT OF FORM I AND THE SUB-TEST SELECTED FROM FORM I BY THE ADJACENT-GROUP METHOD

Test	r	z	1/(n-3)
Form I	.68	.828	.00508
Sub-test from Form I	.58	.663	.00508
	difference	.165	sum .01016

$s_d = \sqrt{.01016} = .101$, $t = .165/.101 = 1.62$, d.f. $= \infty$, $P = .11$.

The difference between the two validity coefficients was not significantly different from zero. The obtained difference could have been caused by chance alone eleven times in one hundred.

These data indicate the difference between the correlation of the external criterion (Form II) and each of the sub-tests (Horst and Group) is not significant. The group method selected a 30-item test from the original test that was not significantly less valid than the original test of 150 items. On the basis of the data of this study, the proposed technique appears to be as good for the selection of items as the more complex method of Horst.[1]

---

[1] Horst, loc. cit.

CHAPTER IV

SUMMARY AND CONCLUSIONS

The major purpose of this study was to present and evaluate a new item-analysis technique applicable in situations where the primary interest is in the discrimination between the members of two or more groups, rather than discriminating between the members within groups. This type of problem occurs frequently in assigning letter grades and in selection work where individuals are to be divided into two groups (that is, those who may be expected to succeed and those who may be expected to fail in a particular situation). The adjacent-groups technique of item analysis resulted in a maximized ratio of between-groups variance to total variance.

The procedure for computing the adjacent-group technique validity indices was presented. It was assumed that the test-score distribution should be the one best fitting the need of the particular situation where the test is to be used. In the case of two groups, the most discriminating test was found to be the one that yields a score distribution with a point of partition at the abscissa of the minimal ordinate between the two group modes. A theoretical examination of the score distribution showed that for any test, a non-overlapping bimodal distribution may be obtained in selecting a sufficient number of items. In the theoretical comparison of the bimodal test-score distribution with a normal test-score distribution, it was demonstrated that the bimodal score distribution resulted in fewer chance errors than the normal distribution. In the case of more than two groups, the distribution should have points of par-

tition at the abscissa of the minimal ordinate between any two adjacent-group modes.

The empirical findings of this study must be interpreted in terms of the small number of cases involved in the data. The selected test items tended to be stable under cross-validation. The empirical studies on the error of measurement of the multimodal test-score distribution showed this error to be minimal at the points of partition separating two adjacent groups; the tendency seemed to be that the error of measurement approached zero at the points of partition.

Since test items discriminating perfectly between two adjacent groups are difficult to obtain, it is quite apparent that the adjacent-groups technique for the selection of items is more feasible in situations where a large source of test items is available. However, the adjacent-group technique can also be applied in situations where intra-group comparisons are to be made and the source of items is limited. The empirical findings showed that the adjacent-group technique was as satisfactory as Horst's more laborious technique of maximizing function in the selection of the most valid items in terms of an external criterion. The adjacent-group technique was found to be superior also to the technique of Flanagan.

If future studies support these findings, it would seem desirable to assume a distribution of scores best suited to the purpose of a test and to select items which will tend to yield this distribution, instead of assuming the usual normal distribution of scores in all situations and selecting items which will tend to yield the normal distribution.

Experience has shown that better prediction is attainable in the tails of a normal distribution than in the center of the distribution;

this could be a possible indication that the normal distribution fails to accurately differentiate between the individuals in the center of the distribution. A multimodal type of test score distribution should increase the efficiency of prediction for the middle group; the overall efficiency of prediction would be increased.

LINDQUIST, E. F., and W. W. Cook, "Experimental Procedures in Test Evaluation," Journal of Experimental Education, 1:163-185, 1933.

LONG, J. A., and P. Sandiford, The Validation of Test Items. Toronto: Department of Educational Research, University of Toronto, Bulletin No. 3, 1935.

MERRILL, W. W., "Sampling Theory in Item Analysis," Psychometrika, 2:215-23, 1937.

RICHARDSON, M. W. and D. C. Adkins, "A Rapid Method of Selecting Test Items," Journal of Educational Psychology, 29:547-52, 1938.

SWINEFORD, F., "Validity of Test Items," Journal of Educational Psychology, 27:68-78, 1936.

VERNON, P. E., "Indices of Item Consistency and Validity," British Journal of Psychology, Statistical Section, 1:152-66, 1948.

VOTAW, D. F., "Graphical Determination of Probable Error in Validation of Test Items," Journal of Educational Psychology, 26:682-86, 1935.

ZUBIN, J., "The Method of Internal Consistency for Selecting Test Items," Journal of Educational Psychology, 25:345-56, 1934.