A METHOD FOR SELECTING COMBINATIONS OF TESTS AND DETERMINING THEIR BEST "CUT-OFF POINTS" TO YIELD A DICHOTOMY MOST LIKE A CATEGORICAL CRITERION

by

RAYMOND FRANZEN

with an
Appendix

Solution of the Selection of the Best Combinations of Dichotomous Arrangements to Distinguish a Categorical Criterion

by

Paul F. Lazarsfeld and Raymond Franzen

A report on a statistical method developed under an grant-in-aid from the Committee on Selection and Training of Aircraft Pilots of the National Research Council, from funds provided by the Civil Aeronautics Administration.

March 1943

CIVIL AERONAUTICS ADMINISTRATION
Division of Research
Report No. 12
Washington, D. C.

National Research Council

Committee on Selection and Training of Aircraft Pilots

Executive Subcommittee

C. W. Bray	J. C. Flanagan
D. R. Brimhall	H. M. Johnson
L. A. Carmichael	W. R. Miles
J. W. Dunlap	G. R. Wendt

M. S. Viteles, Chairman

LETTER OF TRANSMITTAL

NATIONAL RESEARCH COUNCIL

2101 Constitution Avenue, Washington, D. C.
Division of Anthropology and Psychology

Committee on Selection and Training of Aircraft Pilots

March 26, 1943

Dr. Dean R. Brimhall
Director of Research
Civil Aeronautics Administration
Washington, D. C.

Dear Dr. Brimhall:

The attached report entitled A Method for Selecting Combinations of Tests and Determining Their Best "Cut-off Points" to Yield a Dichotomy Most Like a Categorical Criterion, by Raymond Franzen, is submitted by the Committee on Selection and Training of Aircraft Pilots with the recommendation that it be included in the series of technical reports published by the Division of Research, Civil Aeronautics Administration.

The report is not concerned primarily with experimental findings, but with a new method for the analysis of results obtained in investigations on the selection of aviation personnel. To date, this method has been applied in the treatment of data gathered in an investigation conducted by the Committee on Selection and Training of Aircraft Pilots, at the Naval Air Station, Pensacola, Florida, in cooperation with the United States Navy, with funds provided by the Civil Aeronautics Administration.

Frequent reference to the statistical technique will be found in later reports dealing with the analysis of data obtained in the examination of naval aviators at Pensacola. The value of this method, in combination with correlational and other techniques, is being further investigated in connection with the analysis of the findings of the Standard Testing Program and of the Midwest and Boston Projects sponsored by the Committee.

In addition to the main report by Dr. Franzen, there is presented, in Appendix "A", a simplified revision of the method developed by Paul F. Lazarsfeld, in collaboration with Dr. Franzen, which provides a short-cut computational procedure for the solution of problems involving the selection of the best combination of dichotomous arrangements to distinguish a categorical criterion.

In presenting this report attention should again be drawn to the fact that it is devoted to the description of a new method which is to be considered in comparison with other techniques that are available for the treatment of dichotomous material. There has been much discussion within the Committee concerning the advantages of this as compared with other methods, and there is no intention, in presenting this report, to imply that the method should be substituted for the methods more commonly used in experimental studies.

Very truly yours,

Morris S. Viteles, Chairman
Committee on Selection and
Training of Aircraft Pilots
National Research Council

MSV:rm

TABLE OF CONTENTS

A METHOD FOR SELECTING COMBINATIONS OF TESTS AND DETERMINING THEIR BEST "CUT-OFF POINTS" TO YIELD A DICHOTOMY MOST LIKE A CATEGORICAL CRITERION

Introduction.

The present paper provides a new method of combining test scores into a meaningful predictive battery. The particular merit in this method lies in the fact that it permits recovery of useful predictive material from distributions of test scores which, if analyzed by the more common techniques used in test construction, would be discarded. It may be employed without the usual assumptions and problems involved when the conventional methods of test-selection are employed in the construction of a test-battery. Presented below are (a) a general statement of the type of test situation in which this method is useful and more generally applicable than the more common methods of analysis, and (b) a specific example making use of the method and some brief comparisons with other techniques.

The aim underlying the usual techniques of test selection is that of finding a test which is correlated (linearly or curvilinearly) with the criterion. By these methods it is possible to grade or classify the individuals throughout the entire range of test scores. When a dichotomous criterion is used, it is also usually assumed that the variable is actually continuous and that the individuals are normally distributed within the two groups. Some analysis of variance within the groups is therefore needed. In general, three types of statistical techniques are employed in the above type of analysis: (a) Correlational methods — the intercorrelations of the tests themselves and their criterion correlations are studied. Those which show a significant linear or curvilinear correlation with the criterion are combined into a battery using one or more of the various multiple-regression or maximizing techniques. (b) 'Cut-off' procedures — this method, as usually employed, does not determine weights for combining the tests into a battery, but does provide a means by which scores on different tests may be considered together. Here, those tests with significant criterion correlations are employed and a point in the total distribution of each selected at which there seems to be a significant difference in the criterion groups. These tests are then employed singly or in arbitrary combinations (sometimes empirically established) to preselect the subjects. This method of cut-off is, however, no better than erecting a cut-off score based on the average of the standard scores, and still assumes correlated activity throughout the range encompassed by these averages. (c) Analysis of Variance — this technique is often used alone or in combination with the above forms of analysis. Here an attempt is made to answer the question "How can we best weight the tests to get a maximum standard difference in the means?" One such procedure is the Discriminative Function of R. A. Fisher. His method sets up a linear function of the tests and seeks to determine the constants of these functions which will permit the best possible separation of the two criterion groups.

Note here again that assumptions of normality, linearity or curvilinearity, and variance within the criterion groups are present in all the above methods.

It is, however, unnecessary to assume that a test must correlate throughout its entire range or provide a continuous, graded classification of the total population. Further, in those cases where the criterion group is 'fixed' or 'truly' categorical, it is unnecessary to concern ourselves with an estimation

of variance within any single group though complete treatment of the data should include more than one form of analysis.

Suppose that a battery of experimental tests has been given to a sample of individuals who fall into a truly categorical dichotomy, i.e., the status of the criterion is not arbitrary and changing, but is fixed and stable as such. Suppose further, previous analysis has shown that the tests have no dependable linear or curvilinear correlation with this categorical criterion. Now, however, when the total distributions of test scores are re-examined it may be found (as is often the case) that certain areas within the total distribution seem to be related to the criterion while certain others are not, i.e., the scores for one of the criterion groups all tend to cluster in one particular area of the total range, while the scores of the other of the criterion groups are scattered over the remainder of the range, but not included in significant numbers in the area embraced by the first group. This area, if the sample is representative, may then be used in preselecting those individuals who fall into this limited range of scores. This same situation could conceivably be found for a number of tests, each one being, in some degree at least, a good selector in itself. The problem now is to combine these tests (which, although uncorrelated linearly and continuously throughout the entire range, show certain areas of relationship) in such a way as to eliminate a maximum of one of the criterion groups and retain a maximum of the other. (The particular level of efficiency of this elimination is primarily a function of how many of those individuals in the upper ranges we can afford to sacrifice along with those in the lower ranges.)

One other consideration arises here as in the other methods of test construction. Does passing in one or more of the tests compensate for failure in one or more other tests, e.g., does making high scores on psychomotor indices as a whole tend to offset failure in a general intelligence test, so that the individual is actually more like the 'good' criterion group than one would be led to think if only his intelligence scores were considered? It is believed that the method offered in this paper more adequately answers this question of graded compensation than the conventional techniques. Further comparisons with other techniques are, of course, needed before test situations can be defined in which the multiple chi technique is to be employed without recourse to other forms of analysis. In the interest of complete analysis of any data, as many methods as might conceivably yield useful information should be employed.

The proposed method makes use of the chi-squared type of analysis. However, in actual practice, "chi" and not "chi-squared" is used. Briefly: Partial chis, involving passing (at a point chosen statistically) in any number of tests together with failing (at a point chosen statistically) in any other number of tests, are calculated. These partial chis are investigated in all combinations and at all levels of efficiency until one arrangement of successes and failures on the tests (as defined by the cut-offs) is found which presents the best possible prediction of the categorical criterion. This latter combination of partials is called the 'multiple chi'.[1] It should be noted here that multiple, not alternative, use of the partial chis is intended.

---

[1] Full descriptions of terms and techniques mentioned in these introductory pages are presented in the body of the paper.

The prediction of pass and fail in flight training at Pensacola contained elements which suggested the inclusion of this form of analysis as well as the usual procedures. After results of the experimental tests used on the study of the Pensacola cadets had been analyzed by the conventional techniques, the present analysis was undertaken. Areas, such as those mentioned above, which seemed to be related to failure were found in the total range of scores on the tests used in these studies. It was believed also that the criterion of pass-fail in training at Pensacola at the time of this project could be validly considered as 'truly categorical'; at least, the criterion groups were not erected on a-priori considerations of the type of criterion that might be desirable. Criteria in studies of this kind must often be taken just as they stand -- wash-outs vs. passers.

There are, then, three good reasons for viewing the material collected at Pensacola from this point of view as well as from that of correlation throughout the entire range.[2]

1. The psychological factors (all involved in learning) which distinguish the abilities of the upper half of the total distribution are probably different in kind from those that distinguish a 'wash-out' from a successful candidate. In other words, the elements of the distinction are not continuous throughout the whole range of scores.

2. There are many compensatory traits, abilities, and aptitudes which function in the upper half of the distribution, which are probably absent altogether in the lower ranges.

3. The immediate practical problem is to distinguish wash-outs, not to classify or grade success in the upper ranks. This is not meant to imply that this problem does not also exist, but rather that the methods here described are addressed to the former problem and not to this other one. The required method will find ways in which wash-outs respond to test scores which are not shared by the pilots who are successful. But this difference in response does not necessarily imply an average difference. Ideally, when three tests are concerned, this method should cut out that rectangular portion of a cube whose dimensions are the three tests which would place successful candidates at a minimum and wash-outs at a maximum, rather than try to get a linear (triangular) relationship. (App. B and pp. 3-9.)

The following discussion, with examples, provides an empirical way of achieving this result with more tests than three. It is, of course, possible to obtain a generalized formula, but as will be seen in the discussion, there are many by-products of an empirical approach.

Which of the Tests Distinguish the Rejected Group?

A division of the parent population into two groups (one enjoying success and the other rejected as incompetent) provides the criterion for selecting some of an experimental battery of tests which were thought to measure characteristics associated with this criterion. The first step in the analysis is to determine the probability of getting the obtained divergence in one sample (those that were rejected) from corresponding theoretical values in the parent

[2] A short comparison of the efficiency of these two methods when applied to portions of the Pensacola data is given in Appendix B.

population (all the candidates, rejected or not) for each of the tests used.

This is obtained by the use of the chi-squared formula: $\chi^2 = \sum \frac{(o - e)^2}{e}$

o = the "observed" frequency of the criterion group at any score or score interval.

e = the "expected" frequency at the score or score interval.

'e' is obtained by applying the percentage distribution of the scores of the parent population to the N of the criterion group, thereby obtaining a distribution for the group identical with that of the parent population, and against which the "observed" distribution is to be measured. The o-e value is then the actual frequency at any interval minus the frequency that would have occurred if the sample had been distributed exactly as its parent population. This method does not measure the association between variates. It freezes the parent distribution and then measures the probability that the distribution of a given handful will differ as much as this by chance. The p-value of chi-squared used at this point in the analysis states the odds that the wash-outs are a chance withdrawal from the total population.

In all of the examples used in this discussion, the distributions have been made according to intervals of the standard deviation. In other words, the scores of all tests were expressed as standard scores, $\frac{(X - M_X)}{\sigma_X}$, and the frequency distributions are of these standard scores.

In the tables, therefore, individual scores, cut-off points, and distribution intervals on all tests are comparable. The same method of test selection and determination of cut-off points may be used when original scores are retained, but then the values of cut-off points are more difficult to interpret, and without comparable intervals, distributions are more difficult to compare.

Examples A and B below present illustrations of the chi-squared technique as it is employed in this study. Both significant and non-significant samples are chosen:

Example A: Chi-squared for comparing distribution of a sample with its parent distribution (no significant difference):

Standard Deviation Interval	o	% of Parent Pop.	e (% x 36)	o - e	$(o - e)^2$	$\frac{(o-e)^2}{e}$
.7 & above	5	22.2	3.0	-3.0	9.00	1.13
.1 to .6	10	19.7	7.1	2.9	3.41	1.19
0 to -.3	8	20.8	7.5	.5	.25	.03
-.4 to -.6	5	12.5	4.5	.5	.25	.05
-.7 & less	8	24.8	8.9	-.9	.81	.09
Total	36	100.0	36.0	0.0	χ^2 =	2.486

P for 4 degrees of freedom = .73

Example B: Chi-squared for comparing distribution of a sample with its parent distribution (significant difference):

Standard Deviation Interval[3]	o	% of Parent Pop.	e (% x 35)	o - e	$(o - e)^2$	$\frac{(o - e)^2}{e}$
.4 & above	5	38.4	13.4	-8.4	70.56	5.27
.1 to .3	4	11.1	3.9	.1	.01	.002
0 to -.3	5	15.1	5.3	.3	.09	.02
-.4 to -.6	8	12.9	4.5	3.5	12.25	2.72
-.7 to -1.2	8	13.1	4.0	3.4	11.56	2.51
-1.3 & less	5	9.4	3.3	1.7	2.89	.88
Total	35	100.0	35.0		X^2 =	11.40

P for 5 degrees of freedom = .043

It is apparent from an inspection of the 'o' and 'e' frequencies of Example A, that the rejected group is very like the parent distribution. The P of .73 is a measure of this likeness. Seventy-three times out of 100 a difference as large as that obtained between these two distributions would occur when samples of this size were withdrawn from the parent distribution at random. The difference, therefore, is not significant.

It is also apparent from inspection that the 'o' and 'e' frequencies in Example B are more different than in Example A. The rejected group has few in the upper end and many in the lower end of the distribution, when compared with the parent population.

The P of .043 is a measure of the difference. Only four times out of one hundred would groups of this size, selected at random, show as large a difference as this from the "expected" distribution. The difference, therefore, may be attributed to the characteristics of the sample.

At What Point in the Scale Shall Test Failure be Placed?

In order to determine the point of "cut-off" defining "pass" and "fail" two measures are used. First, chi, in order to determine the point of cut-off that yields the most significant difference between the number of the criterion group failed at that point, and the number of the parent population failed at that point, i.e., varying degrees of selective efficiency are not considered. Only the greatest efficiency is tested for by showing the cut-off at which the pass and fail division of the unsuccessful pilots is most different from the

---

[3]Before computation regular intervals are combined so that no interval of the 'o' or the 'e' distribution has less than 3 cases.

same division of all pilots.[4] The formulae for obtaining theoretical values and for chi itself, remain the same as when applied to the distribution, but chi, not chi-squared, is used because there is only one degree of freedom (two classes of each variable).[5] Second, percent of failure in the parent population since the test discrimination must be severe enough to eliminate only a small proportion. As a rough standard of severity, we have assumed that in order to be considered, a cut-off on one test or a combination of cut-offs on a battery of tests used as pre-selectors, must not fail more than twenty percent of those who are retained by experience.

Example C: Presented below is an example of chi for comparing categories of a sample with categories of its parent population (data from example B, page 5). Before combining intervals, the distribution of the criterion group and the expected distribution are calculated as follows:

Standard Deviation Interval	o	e
1.8 & above	-	1.6
1.7 to 1.5	1	1.7
1.4 to 1.0	1	2.8
.9 to .7	-	2.8
.6 to .4	3	4.5
.3 to .1	4	5.9
0	-	1.3
-.1 to -.3	5	4.0
-.4 to -.6	8	4.5
-.7 to -.9	3	2.8
-1.0 to -1.2	5	1.8
-1.3 to -1.7	3	1.7
-1.8 & less	2	1.6
Total =	35	35.0

---

4 Again we are not measuring association between variates, but merely the probability that the difference which we obtain between our rejected group and the parent would occur if the group were selected at random.

5 The probability of a chi (when grouping reduces the degrees of freedom to 1) is interpreted like the probability of a standard score. A chi of 1.0 indicates a probability of .32 that the difference was the result of chance; for a chi of 1.9 the p-value is .05 (.025 at each end of the distribution); etc.

To test for various cut-off points these distributions are cumulated into the following pass and fail categories and their chis calculated as follows:

Cut-off at a score of		o	e	o - e	$(o - e)^2$	$\frac{(o - e)^2}{e}$	chi
-1.3	Pass	30	31.7	1.7	2.89	.091	0.98
	Fail	5	3.3	1.7	2.89	.876	
						.967	
-1.0	Pass	25	29.9	4.9	24.01	.803	2.35
	Fail	10	5.1	4.9	24.01	4.709	
						5.512	
- .7	Pass	22	27.1	5.1	26.01	.960	2.06
	Fail	13	7.9	5.1	26.01	3.292	
						4.252	
- .4	Pass	14	22.6	8.6	73.96	3.273	3.04
	Fail	21	12.4	8.6	73.96	5.965	
						9.238	
- .1	Pass	9	18.6	9.6	92.16	4.955	3.25
	Fail	26	16.4	9.6	92.16	5.620	
						10.575	

These chis indicate that -.4 or -.1 are the "cut-offs" yielding the most significant number of "failures" in the criterion group (that is, the group who were rejected by experience). At -.4, 60 percent of the rejected group and 35 percent of the parent population are failed. A failure point at -.1 nets 74 percent of the rejected group and 47 percent of the parent population.

This one test, therefore, cannot be used alone to select cases having the characteristics of those who were rejected by experience. Although it shows a reliable difference between the distribution of the rejected and of the parent groups and a reliable difference in "pass" and "fail" categories at two cut-off points, both of these points fail too many of the parent distribution.[6]

Since the other tests which showed reliable differences also fail too many at reliable "cut-off points," it becomes necessary in this situation to use more than one test.

How May These Tests be Combined to Establish a Composite Cut-off Point (Index of Pass or Fail)?

By application of the chi methods already discussed to a battery of tests, three tests (for example) have been selected. These three all show significant differences between the distribution of the scores of the group

[6] A chi of approximately 1.90 represents a p-value 0.05.

rejected by experience and of the parent population. (See below.)

Test	P	χ^2	Degrees of Freedom
a (not Test A of previous discussion)	.030	12.41	5
b (Test B of previous discussion)	.043	11.40	5
c	.00028	23.52	5

The cut-off points in certain levels in the distributions of standard scores show the following characteristics:

Cut-off Point	Chi			% of Rejected Group Failed			% of Total Population Failed		
	Test a	Test b	Test c	Test a	Test b	Test c	Test a	Test b	Test c
-1.0	1.35	2.35	0.75	26	29	23	18	15	18
-.7	2.08	2.06	2.47	41	37	40	26	23	23
-.4	2.23	3.04	3.62	56	60	57	37	35	29
-.1	2.66	3.25	3.95	68	74	74	45	47	41

On all of these tests -.4 and -.1 as cut-offs yield significant chis. All, however, select too many of the non-criterion cases at these points.

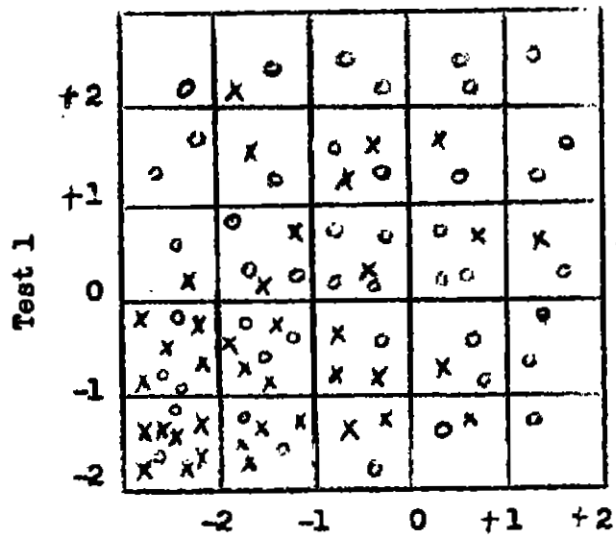
Note that each of the tests alone approximates a good selective device, but it is not known whether failure on all three or on any two and not the other is the best standard for failing the rejected group without failing too many of the retained cases.

If superior performance on one test acts as compensation for the inferior ability demonstrated by another test, then failure on both tests will be a better standard of elimination than failure on either test used alone. When there are three tests, the possibilities for determination of failure are increased. If passing either of two tests compensates for failing the third, then pass-fail-fail and fail-pass-fail, as well as fail-fail-fail standards should be investigated. Satisfactory ability demonstrated by passing one of the three tests may be all that is necessary to compensate for the handicap indicated by failure on the others. If this is the case then failure on all three tests will be required to predict the criterion characteristics.

As part of this problem of the proper combination of pass and fail on each of the three tests is the question of the point on each test below which a score shall be considered as failure. All possible combinations of the three tests at all possible success and failure cut-off points (levels of elimination in the distribution) are first analyzed in order to find the combination of cut-off points which will yield the largest number of failures among those actually rejected by experience accompanied by the smallest number of other failures among those retained by experience.[7]

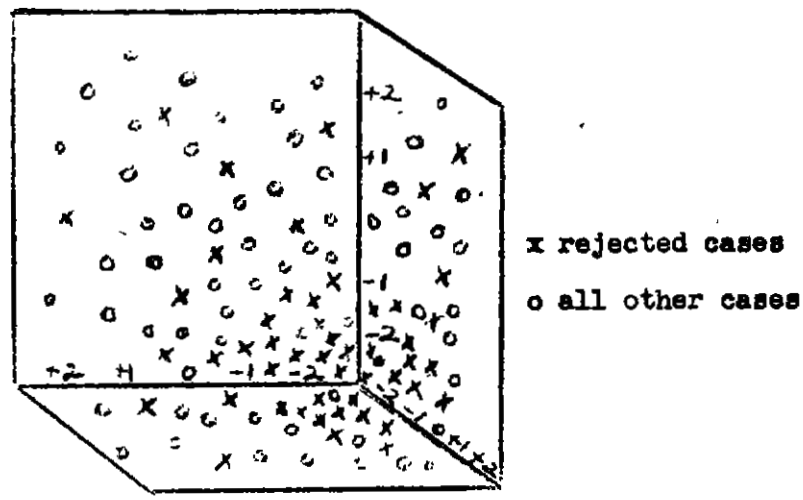
[7]This operation is very laborious. A shorter, more practical method, saving a great deal of the computational labor, appears in Appendix A.

When only two tests are considered, the problem is easily presented in a graphic manner:



What rectangle or square in the lower left-hand corner will enclose the largest number of failures (X's) and the smallest number of passers (circles)?

When three tests are considered a graphic description may be presented by a cube. The X's and circles are suspended in positions determined by their relation to the scales on the three dimensions of the cube. What solid in the corner, which represents the lowest scores of each test, will enclose the largest number of failures (X's) and the smallest number of passers (circles)?



This problem, mathematically, is one of testing the difference between the proportion of pass and fail rejected cases and the proportion of pass and fail retained cases at various definitions of failure. These definitions are at different levels of the distribution on different tests in combination.

How May We Find the Best Definition of Failure?

Up to this point the analysis has dealt entirely with wash-outs versus all pilots (total population). The problem now arises specific to pilot selection of how these tests may be used to distinguish the wash-outs (those eliminated by experience) from those who are retained by experience (passers), i.e., the problem is to locate a point in aptitude (measured by the tests) below which all pilots will be rejected (unsuccessful performance) and above which they will be retained (at least adequate performance).

In order to investigate these tests in combination as they operate to distinguish wash-outs from successful pilots the same chi technique is applied in the form of the multiple chi. This multiple chi is arrived at by an analysis of all of the possible combinations of the partial chis in the test situation.

The chi formula to test for associations between variates in a contingency table is as follows: Let the four values (a, b, c, and d) in the two categories (rejected and retained) be indicated in a two-by-two contingency table as follows:

	Failed by Test	Not Failed by Test	Total
Rejected by experience	a	b	a + b
Retained by experience	c	d	c + d
Total	a + c	b + d	a + b + c + d

In this situation the formula for chi, $\sqrt{\sum \frac{(o-e)^2}{e}}$ becomes:

$$\sqrt{\frac{(a+b+c+d)\ (ad-bc)^2}{(a+b)\ (c+d)\ (a+c)\ (b+d)}}$$

Example of chi to test for associations between variates in a contingency table:

	Failed by Test	Not Failed by Test	Total
Rejected by experience	3	31	34
Not rejected by experience	42	255	297
Total	45	286	331

$$\text{Chi} = \sqrt{\frac{331\ (765 - 1302)^2}{(34)\ (297)\ (286)\ (45)}} = .86$$

Since there are scores for 34 criterion cases and for 297 non-criterion cases on all three tests, (a + b) (c + d) and (a + b + c + d) have the same value throughout all combinations.

The table on page 12 gives pass and fail categories and chis obtained for all possible combinations of the three tests at the -1.0 and less, -.7 and less, -.4 and less, and -.1 and less, levels of failure.

A chi is given a negative value if the difference it measures is in the unexpected direction. A negative chi then indicates that the observed value of cases failed by test and rejected by criterion is less than the expected value, i.e., there are fewer wash-outs and more of the group retained by experience selected by the given pass-fail level than would be expected from the total distribution of pilots.

It is also possible that the same level of failure should not be used on all three tests. In the materials being used as examples, for instance, chis were computed for all seven combinations when the failure level was -.1 on two of the tests, but was -1.0 on the other, for all the seven combinations with the same failure level on two of the tests, when it was -.7 on the other and also when it was -.4 on the other. There were then 63 such chis. All of these must be considered because being the best cut-off when a test is used alone does not necessarily mean it is best for that test in combination with others. In the interests of simplicity, however, this discussion will be limited to the data in the table where the same level of failure was used for all three tests in their pass and fail combinations.

The chis presented in the preceding table are in reality partial chis,[8] for by combining the fail categories of different combinations we may obtain any of the following composite determinations of failure at any cut-off point:

Call:

a.bc (failure on a but pass on b & c)	1
b.ac (failure on b but pass on a & c)	2
c.ab (failure on c but pass on a & b)	3
ab.c (failure on a & b but pass on c)	4
ac.b (failure on a & c but pass on b)	5
bc.a (failure on b & c but pass on a)	6
abc (failure on a & b & c)	7

A chi then, is obtained for:	by combining the failure of:
Failure on a without consideration of b & c	1, 4, 5, & 7
Failure on b without consideration of a & c	2, 4, 6, & 7
Failure on c without consideration of a & b	3, 5, 6, & 7
Failure on at least one of a or b without consideration of c	1, 2, 4, 5, 6, & 7
Failure on both a & b without consideration of c	4 & 7
Failure on at least one of a or c without consideration of b'	1, 3, 4, 5, 6, & 7
Failure on both a & c without consideration of b	5 & 7
Failure on at least one of b or c without consideration of a	2, 3, 4, 5, 6, & 7
Failure on both b & c without consideration of a	6 & 7
Failure on at least one of a or b or c	1, 2, 3, 4, 5, 6, & 7
Failure on all three of a & b & c	7

[8] Note - Multiple rather than 'alternative' use of these partial chis is intended. In the actual computational process it is unnecessary to carry these partial measures through to the final chi. A methematical shortcut by P. F. Lazarsfeld and Raymond Franzen is presented in Appendix A.

COMPARISON OF PASS AND FAIL CATEGORIES OF REJECTED GROUP AND OF NON-REJECTED GROUP IN ALL COMBINATIONS OF TESTS AT VARIOUS LEVELS OF FAILURE

Combination of tests	Failure Level on All Three Tests: -1.0 & less		-.7 & less		-.4 & less		-.1 & less	
	Not Failed	Failed	Not Failed	Failed	Not Failed	Failed	Not Failed	Failed
(1) a.bc[9]								
Rejected cases	2	32	4	30	3	31	2	32
Retained cases	31	266	35	262	42	255	33	264
Chi	-.84		-.003		-.86		-.94	
(2) b.ac								
Rejected cases	6	28	5	29	5	29	5	29
Retained cases	19	278	29	268	33	264	32	265
Chi	2.35		.90		.62		.69	
(3) c.ab								
Rejected cases	3	31	3	31	2	32	2	32
Retained cases	29	268	30	267	23	274	26	271
Chi	-.18		-.24		-.39		-.57	
(4) ab.c[10]								
Rejected cases	2	32	3	31	2	32	3	31
Retained cases	6	291	13	284	27	270	36	261
Chi	1.39		1.15		-.62		-.57	
(5) ac.b								
Rejected cases	4	30	6	28	6	28	5	29
Retained cases	10	287	15	282	17	280	21	276
Chi	2.31		2.85		2.59		1.57	
(6) bc.a								
Rejected cases	2	32	3	31	4	30	4	30
Retained cases	11	286	13	284	19	278	29	268
Chi	.62		1.15		1.17		.37	
(7) abc								
Rejected cases	-	34	2	32	8	26	13	21
Retained cases	4	293	7	290	20	277	38	259
Chi	-.68		1.20		3.33		3.89	

[9] a.bc means failure is defined as below failure level on Test a but above failure level on Tests b and c.

[10] ab.c means failure is defined as below failure level on Tests a and b but above failure level on c.

Thus at -.1 cut-off the chi of failure on a, without consideration of b or c, is obtained from the following table:

	Failures	Passing	Total	Chi[11]
rejected at 1, 4, 5, & 7	23	11	34	2.72
retained at 1, 4, 5, & 7	128	169	297	

Failure on b, without consideration of a or c, is obtained from:

	Failures	Passing	Total	Chi[11]
rejected at 2, 4, 6, & 7	25	9	34	3.10
retained at 2, 4, 6, & 7	135	162	297	

Similarly, the following chis are computed:

	Chi
Failure on c, without consideration of a or b.	3.61
Failure on at least one of a or b, without consideration of c.	3.57
Failure on both a & b, without consideration of c.	2.75
Failure on at least one of a or c, without consideration of b.	2.73
Failure on both a & c, without consideration of b.	4.32
Failure on at least one of b or c, without consideration of a.	3.79
Failure on both b & c, without consideration of a.	3.48
Failure on at least one of a or b or c.	3.53
Failure on all three of a & b & c (the chi appearing on the table).	3.89

Most of these chis are high enough to reflect a very significant difference between the two groups in their proportion of pass and fail. They are, of course, much higher (with the exception of abc which is the same) than those under -.1 on the table because there is compensation between the elements. When passing on one or more of the tests is part of the composite standard of failure, then we obtain lower chis. It is apparent that either failure on all three, or failure on two with no consideration of the score on the third, will provide the best determinations of rejection.

---

[11]Chis on failure on a, without consideration of b or c; failure on b, without consideration of a or c; and failure on c, without consideration of a or b; may be compared with the cumulative two-place chis for -.1 cut-off given on p. 8. They are slightly different because the formula used there was one comparing a sample with its parent population.

The question remains as to which of these composites will fail the smallest portion of the retained group while failing a large enough proportion of the rejected population (see following table).

	Chi	% of Rejected Group Failed	% of Retained Group Failed
Below -.1 on a & c without consideration of b	4.32	53	20
Below -.1 on a & b & c	3.89	38	13
Below -.1 on either b or c without consideration of a	3.79	94	61
Below -.1 on c without consideration of a & b	3.61	71	38
Below -.1 on either a or b without consideration of c	3.57	94	64
Below -.1 on a or b or c[12]	3.53	100	72
Below -.1 on b & c without consideration of a	3.48	50	23
Below -.1 on b without consideration of a or c	3.10	74	45
Below -.1 on a & b without consideration of c	2.75	47	25
Below -.1 on either a or c without consideration of b	2.73	85	62
Below -.1 on a without consideration of b & c	2.72	68	43

The purposes of the particular problem for which the tests have been selected will determine which of the composites to adopt. If the need is to fail as many as possible of the cases who will be rejected by experience, even at the sacrifice of others, then test c alone would probably be selected. If the problem is the more usual one of failing a maximum of the rejected cases together with a minimum of those retained, then failure on both a and c is the most acceptable standard. If, on the other extreme, a minimum sacrifice of cases is essential, then failure on all three tests should be required.

For purposes of explanation, the analysis of composites has been limited to the failure level of -.1. In the analysis from which this example was drawn, the same procedure was followed for lower failure values and also for differing failure values on the tests being examined.

One important advantage gained from this form of analysis is information in reference to the compensatory relation among tests. The calculations tell you whether failure in one or two of the tests is as highly related to the criterion as is the failure on all of them together. Assume now that two of the tests involved are psychomotor tests and that the other one is a general intelligence test. Knowing that these tests are all related to the probability of rejection by experience, would not reveal whether failure in general

---

[12] This situation is analogous to the 'Hurdle Method' of cut-offs.

intelligence is compensated by success in the psychomotor examinations, or vice versa. When the analysis described above is applied it might be found that such compensation does exist. It would then be known that being low in psychomotor ability, but above average intelligence, does not predict rejection nearly as well as being low in both traits. Likewise it would be known that being low in intelligence, but above average in psychomotor ability does not predict rejection as well as being low in both. Such interpretations of compensation are possible for any number or any combination of tests. If the analysis is done step by step as outlined above, direct evidence of the manner of compensation is developed.

Summary of Conditions in Which The Multiple Chi Technique Is More Generally Applicable Than The Usual Procedures.

Briefly: The use of the multiple chi technique of constructing a battery of selection tests is warranted if the following conditions or circumstances are found.[13]

(a) When the criterion is 'truly' categorical, i.e., is not erected on *a priori* considerations.

(b) When, therefore, we are not concerned with the variance in any single criterion group within the categorical breaks.

(c) When we wish only to distinguish one group from the other even if there is no average difference.

(d) When we do not wish to grade, scale, or classify degrees of success, failure, performance, etc.

(e) When the distributions of scores show certain areas within their total distributions that are related to the criterion and other areas that are not, yet which do not correlate (show consistent and continuous relation with the criterion throughout the entire range of scores).

(f) Where graded compensation between test scores is more important than reinforcement.

---

[13] These conditions or circumstances are not mutually exclusive.

APPENDICES

A. Solution of the Selection of the Best Combinations of Dichotomous Arrangements to Distinguish a Categorical Criterion, by P. F. Lazarsfeld and Raymond Franzen.

B. Comparisons of 'Multiple Chi' and Multiple Regression Techniques.

APPENDIX A

Solution of the Selection of the Best Combinations of Dichotomous Arrangements to Distinguish a Categorical Criterion

Paul F. Lazarsfeld has contributed a simple and direct solution of this problem. Instead of using partial chis as was done in this report, he has suggested using partial deltas. The advantage is in economy of computation. The delta indicates the degree of relationship (value of the chi) during the chi computation and thus allows for selection of combinations without computing each separate chi. Furthermore the deltas are additive and therefore themselves produce the value from which the desired combined chi is obtained. Computing many chis having the same N by means of the deltas also allows for a simple tabular arrangement of the data which facilitates computation.

The following formulae define delta, using a, b, c, and d to designate the four cells in a 2 by 2 contingency table:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc = \triangle$$

Actual minus theoretical frequency within a cell $= d = a - \frac{(a+c)(a+b)}{a+b+c+d} = \frac{ad - bc}{N} = \frac{\triangle}{N}$

The frequencies for the cells are, of course, obtained from sorting. As an example of the use of delta in the problem of selecting combinations of tests that will predict flying inaptitude (being washed-out), let us designate passing on a test with a given criterion by plus, and failing on that same test with that same criterion by minus. Thus, if below average is to be considered failing, then plus on Test 1 would mean that the case was above average and minus on Test 1 would mean that the case was below average. Plus on 1, 2, and 3 would mean passing on all three tests. Plus on 1 and 2 and minus on 3 would mean passing 1 and 2 and failing 3. After all combinations of pass and fail on the three tests have been tabulated we will have an arrangement such as appears on page 20. The top row of numbers refers to the number of wash-outs who have each arrangement of pass and fail on the three tests. The lower row of numbers refers to numbers of retained pilots who have each arrangement of pass and fail. For example, two wash-outs and 26 retained pilots passed (were above average on) Tests 1 and 2 and failed (below average) Test 3.

Test	Combinations of Pass (+) and Fail (-) on 3 tests.								Total
1	+	+	+	+	-	-	-	-	
2	+	+	-	-	+	+	-	-	
3	+	-	+	-	+	-	+	-	
Wash-outs	0	2	5	4	2	5	3	13	34
Retained	82	26	32	29	33	21	36	38	297

We may now consider a simple way of expressing each of these pass-fail arrangements as a delta. The figures of the top row are the a cells of the various failure arrangements of the tests and those of the bottom row are the c cells.

We want to get ad - bc, but we may avoid using the d and c values of each arrangement by the simple transformation: ad - bc = a (c+d) - c (a+b), since the (c + d) and (a + b) values are the same for each arrangement.

In the above table (c + d) is all retained cases, or 297, and (a + b) is all wash-outs, or 34.

Now to obtain a (c + d) minus c (a + b) we need only to multiply the top row of figures by 297 and the bottom row of figures by 34 and then subtract (see below).

Test	Combinations of Pass (+) and Fail (-) on 3 Tests							
1	+	+	+	+	-	-	-	-
2	+	+	-	-	+	+	-	-
3	+	-	+	-	+	-	+	-
a (c + d)	0	594	1485	1188	594	1485	891	3861
c (a + b)	2788	844	1088	986	1122	714	1224	1292
Diff.	-2788	-290	397	202	-528	771	-333	2569

We may now pick any combination of pass and fail arrangements we wish and determine its delta. All deltas are additive. If we wish to know the delta of failure on Tests 2 and 3, irrespective of Test 1, we add all arrangements in which 2 and 3 show failure.

If we wish to translate any combination of these deltas into chi, we use the following formula:

$$\text{Chi-squared} = \frac{\triangle^2 \; N}{(a+b)\;(a+c)\;(b+c)\;(b+d)}$$

The a, b, c, and d appearing in this formula are, of course, the a, b, c, and d of the combination for which a chi-squared is computed.

This method of computation has three very important values:

1. It allows us, without computing the chis, to find the partial elements which alone or together make the greatest distinction between the categories of the criterion.

2. It shows analytically, prior to obtaining the final chi values, the compensatory relations that exist among the tests.

3. It has great economy in computing-labor.

APPENDIX B

The following question has been asked of the present method: "Can we arrive at better results with these (Pensacola) materials by using the total range of the tests in correlation with our accepted categorical criterion?" If the multiple regression technique is used a triangular cut-off (Fig. 1) is obtained, i.e., those people in the lined areas are rejected. If the multiple chi is used we get a square cut-off (Fig. 2), rejecting those in the lined regions. This is true of a cube (using three tests) as well as of a two-dimensional surface (two tests).

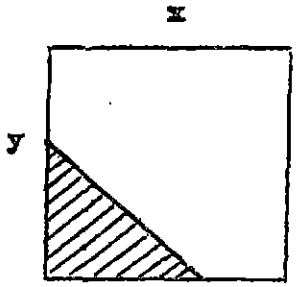


Fig. 1
Cut-off by
Multiple Regression

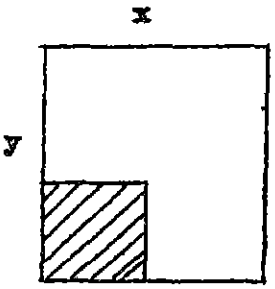


Fig. 2
Cut-off by
Multiple Chi

To answer the above question, it is necessary to assume that correlation with the categorical criterion exists throughout the entire range. Test scores on the Otis Intelligence Test, the Washburn, and the Two-Hand Coordinator were used in this comparison at two levels of selection -- namely, the levels at which 25% and 40% of the wash-outs were selected by the cut-off technique. These two levels were chosen (a) because less than 25% reduces the number of wash-outs too much while over 40% obviously will include too many of those normally retained by experience, and (b) because these are among the most efficient points of selection when either method is used.

The multiple regression technique was employed as follows:

1. Bi-serial correlations between tests and the criterion (wash-out and retained) were calculated.
2. Intercorrelations of all tests were likewise calculated (Pearson).
3. Betas in the multiple regression were employed to obtain theoretical standard score values for the dependent variable.
4. Counts of wash-outs and retained pilots falling below a point in this multiple prediction were then made.

The multiple chi was used as explained in the text.

The accompanying table bears out the assumption that the multiple chi would yield better results than the multiple regression technique in a test situation such as that encountered at Pensacola when the two selected levels used here are employed.

PERCENT OF WASH-OUTS AND OF REMAINING CADETS REJECTED AT VARIOUS CUT-OFFS BY MULTIPLE REGRESSION AND BY MULTIPLE CHI

Otis Intelligence (2), Two-Hand Coordination (4) and Washburn Serial Reaction (6).

PART I DATA

Failure level when determined by:	Wash-outs (N=34) % Rejected	Remaining Cadets (N=297) % Rejected
Chi at -.4 sigma on tests 2 & 6	41	12
*R at theoretical $\bar{y}$ = -.4 sigma	41	18
Chi at -.4 sigma on tests 2, 4 & 6	24	7
*R at theoretical $\bar{y}$ = -.57 sigma	24	13

PART II DATA

Failure level when determined by:	Wash-outs (N=47) % Rejected	Remaining Cadets (N=203) % Rejected
Chi at -.4 on tests 4 & 6	40	12
*R at theoretical $\bar{y}$ = -.5 sigma	40	17
Chi at -.4 sigma on tests 2, 4 & 6	26	06
*R at theoretical $\bar{y}$ = -.75 sigma	26	09

CORRELATION MEASURES

	PART I (N=297)	PART II (N=250)
Bi-serial r's		
r_{12}	.332	.406
r_{14}	.286	.468
r_{16}	.350	.275
Inter-r's (Pearson)		
r_{24}	.245	.220
r_{26}	.290	.137
r_{46}	.297	.214
$R_{1.246}$	.451	.582
Beta 1.46	.224	.305
" 4.16	.160	.369
" 6.14	.238	.155

* - That is, using multiple regression