DETAILED STATISTICAL ANALYSIS OF DATA OBTAINED IN THE PENSACOLA STUDY OF NAVAL AVIATORS

by

Raymond Fransen

and

Ross A. McFarland

A report on research conducted at the Naval Air Station, Pensacola, Florida, in cooperation with the Bureau of Aeronautics of the U. S. Navy and the Division of Research, Graduate School of Business Administration, Harvard University, by means of a 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.

January 1945

CIVIL AERONAUTICS ADMINISTRATION
Division of Research
Report No. 41
Washington, D. C.

National Research Council

Committee on Selection and Training of Aircraft Pilots

Executive Subcommittee

M. S. Viteles, Chairman

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

National Research Council

1945

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

January 19, 1945

Dr. Dean R. Brimhall
Director of Research
Civil Aeronautics Administration
Washington 25, D. C.

Dear Dr. Brimhall:

Attached is a report entitled Detailed Statistical Analysis of Data Obtained in the Pensacola Study of Naval Aviators, by Raymond Franzen and Ross A. McFarland. This report is submitted by the Committee on Selection and Training of Aircraft Pilots with the recommendation that it be included in the technical reports issued by the Division of Research, Civil Aeronautics Administration.

In the fall of 1939 the Committee on Selection and Training of Aircraft Pilots undertook a major study at the Pensacola Naval Air Station, Pensacola, Florida, in the selection and training of aircraft pilots in cooperation with the U. S. Navy using funds provided by the Civil Aeronautics Administration. In this investigation, known as the Pensacola Study, a large variety of psychological and physiological tests were administered to successive classes of aviation cadets and to a group of pilot instructors.

The results of this investigation have been summarized in C.A.A. Division of Research Report No. 38. The present report is concerned with the application of highly specialized statistical techniques considered useful in the detailed analysis of the significance of the predictors employed in the investigation. An earlier report, C.A.A. Division of Research Report No. 12, describes in greater detail the multiple chi technique used in the analysis of the data discussed in the present report.

Cordially yours,

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

MSV:rm

CONTENTS

SUMMARY

This report presents a detailed discussion of the statistical analysis of psychological and physiological test data obtained during a study of naval aviation cadets conducted from July, 1940 to May, 1941 at the Naval Air Station, Pensacola, Florida. Two samples of subjects were employed in this investigation, designated as Part I and Part II. The subjects comprising Part I were officers and college graduates, while those comprising Part II were high school graduates and those who had had an additional two years of college work. All of the trainees had been given 10 hours of dual flight instruction and had soloed before entering Pensacola for further training.

Section I of this report describes the analysis of Part I data on the entire battery of 12 psychological and psychomotor tests and 21 physiological tests. In the analysis of Part II data, described in Section II, those tests which showed significant association with success in flight training in Part I of the study were subjected to further investigation. Since the samples differed for the two sections of the study, the analysis on Part II was carried through separately and compared with the results obtained from the Part I data.

In order to select the best battery of tests from all of the measures employed the following procedures were used:

1. The reliability of test samples was examined to determine if the samples could be regarded as homogeneous. The distribution of scores for each population tested was divided into halves and chi-squared tests applied to measure the significance of a divergence between two samples drawn from the same universe. High chi-squared P-values (indicating homogeneity of sample) were obtained for all tests on Part I, except the Two-Hand Coordination Test, Tilt Chair Test, Thorndike-Kelley Athletic Achievement Test, Tidal Air/Body Surface, and the Ophthalmograph Test. The samples of these five tests, with P-values of .06 or less, were suspect, since such differences, approaching significance, reduced measurably the probability that the differences between the distributions of the halves were chance ones.

A similar analysis was made of the Part II data of five tests (Eye-Hand Coordination Test, Otis Test of Mental Ability, Two-Hand Coordination Test, Mashburn Serial Action Test, and the Minnesota Paper Form Board). All of the P-values were high when the halves of each of these measures were compared. A reversal was obtained of the comparison of the Two-Hand Coordination Test, but this difference was attributed to the lack of calibration of the instrument during the testing of the first group.

2. In Section I intercorrelations among the nine psychological and psychomotor tests showed that both linear and curvilinear relationships among the tests were low. The scores on the Mashburn Serial Action Test correlated most highly with the scores on the other tests, the stas being close to .30 with every test, except Perception of Change in Position and

the Athletic Achievement Test. The highest single relationship was between the Otis and the Minnesota Paper Form Board with an r of 0.36 and an eta of 0.39. It was suggested that in as far as these tests were reliable, these measures, with the possible exception of the Mashburn, evaluated different characteristics in pilot selection.

3. The next step was the selection of those tests most predictive of success in flight training. To select such tests the distributions of test scores of washouts for inaptitude, cadets dropped from training for reasons other than aptitude, and board appearances were compared with the total distribution of test scores of all cadets. The comparison of the group of washouts for inaptitude with the total population by means of the chi-squared test showed that for the Part I sample the P-values for the Otis Test of Mental Ability and the Two-Hand Coordination Test were .03 and .04 respectively while the P-value for the Mashburn on the same sample fell far below the 1% level. For the Part II sample, however, the P-values for the Otis and the Two-Hand Coordination Tests fell far below the 1% level of significance, while the P-value for the Mashburn Test rose to .03. Since in Part II the P-values of two additional tests (Eye-Hand Coordination Test and the Minnesota Paper Form Board) approached the 1% level of significance, it was decided to include these two measures of the Part II analysis along with the selected tests, i.e., the Otis, Two-Hand Coordination Test, and the Mashburn Serial Action Test.

4. Following the selection of those measures which attained or approached reliable differentiation of the washouts from the total population (Otis Test, Two-Hand Coordination Test, Mashburn Test in Section I and Otis Test, Two-Hand Coordination Test, Mashburn Test, Eye-Hand Coordination Test, and Minnesota Paper Form Board in Section II) attention was turned to the problem of locating in the distribution of scores for each of the tests of Parts I and II a cut-off point which would best separate the potential washouts from the potential successes. This was determined by computing chis for failure levels at cut-off points placed at every three-tenths of a standard deviation, and separately for halves of each test. On the basis of a significant chi and of a standard of efficiency requiring that the cut-off point must fail at least 50% of the failures and not more than 20% of those retained, the following tests of Part I were found to meet the criterion sufficiently well to warrant their possible use as predictive tests: Otis Test at the standard score level of -.7, Two-Hand Coordination Test, and the Mashburn Test both at the standard score level of -.4. A similar analysis of Part II indicated that the most efficient selections based on these data were those by the Otis Test at the -1.0 standard score level and the Mashburn Test at the -.7 standard score level.

5. Finally, although each of the tests in Section I and Section II were found to approximate good selection devices, they all failed too many successful pilots at points where they eliminated a sufficient number of washouts. It was necessary to go a step further and, by means of multiple chi, determine for each sample how the tests operated in different combinations at various failure levels. The most promising combination on the Part I data was found to be the Otis Test of Mental Ability (below a standard score level of -.4) and the Mashburn Serial Action Test (below a standard score level of -.1). This combination came nearest to meeting the 50-20% failure requirement.

Since the patterns of failure on the Otis, the Two-Hand, and the Mashburn were quite similar in the two samples, Part I and Part II were combined and various rejection levels for various combinations of these three tests were studied. The Otis Test and the Mashburn Test were found to be efficient predictors when -.1 sigma was the cut-off point for the Mashburn and -.4 the cut-off point for the Otis Test. These levels rejected two-fifths of the washouts and one-fifth of the passers.

An advantage gained from this type of analysis indicates whether failure in one test or two of the tests is as highly related to the criterion as is failure on all of them together. Empirically, we do not know if failure in psychomotor tests is compensated by success in general intelligence, or vice versa. The analysis described in this report showed that compensation did exist, i.e., poor performance in psychomotor tests, but better than average general intelligence did not predict rejections nearly as well as being low in both traits, and conversely, being low in general intelligence, but scoring above average in psychomotor ability, did not predict rejections as well as being low in both.

DETAILED STATISTICAL ANALYSIS OF DATA OBTAINED IN THE PENSACOLA STUDY OF NAVAL AVIATORS

INTRODUCTION

An earlier report[1] in this series has presented a general summary of the studies in aviation psychology undertaken at the Pensacola Naval Air Station under the auspices of the Committee on Selection and Training of Aircraft Pilots. That report presents the general objectives of the study, a description of the experimental design, the tests and measures employed, and the major results. It is the purpose of the present report to present in full detail an extensive statistical treatment of the data gathered in the Pensacola studies.

It will be recalled that the Pensacola studies were conducted with two separate populations which were different in two characteristics, namely, amount of education and in experimental conditions. In the summary report, these different samples were referred to as Parts I and II.[2] The same division is maintained in the present report; all statistical analyses being carried out independently on the two samples. Analyses of the Part I data are presented in Section I of the report. Section II presents the results of the treatment of the Part II data and a comparison of the two groups of subjects.

TESTS AND MEASURES ANALYZED[3]

The tests and measures indicated below are analyzed for their predictive significance in one or both parts of this report. In most instances in the tables the tests will be referred to by number.

Test Number	Name of Test
1	Eye-Hand Coordination*
2	Otis Test of Mental Ability*
3	Ataxiameter - amount of body sway in mm.**
4	Two-Hand Coordination Test*
5	Perception of Change in Position - Tilt Chair*
6	Mashburn Serial Action Test*
7	McDougall Dotting Test*

---

[1] McFarland, Ross A. and Franzen, Raymond. The Pensacola study of naval aviators, Final summary report. Washington, D. C.: Civil Aeronautics Administration Division of Research, Report No. 38, November 1944.

[2] Ibid.

[3] A description of these measures and their scoring systems is presented in the earlier report. Ibid.

Test Number	Name of Test (cont.)
8	Minn. Paper Form Board Test*
9	Thorndike-Kelley Athletic Achievement*
10	Diastolic Blood Pressure - reclining**
11	Systolic Blood Pressure - reclining**
12	Pulse Rate - reclining**
13	Schneider Index#
14	Tidal Air/Body Surface#
15	Ophthalmograph - fixations per line**
16	Basal Metabolic Rate#
17	Cold Pressor - greatest diastolic change toward a positive direction**
18	Cold Pressor - greatest systolic change toward a positive direction**
19	Tilt Table (a) Pulse Pressure Change** (b) Pulse Rate Change** (c) Smallest Pulse Pressure** (d) Time Interval to Smallest Pulse Pressure**
24	Startle - duration of somatic tremor**
25	Startle - amplitude of somatic tremor**
26	Startle - maximum change in rate in a positive direction**
27	Cattell Continuous Reaction Test - Part I#
28	Cattell Continuous Reaction Test - Part II#
29	Vital Capacity/Body Surface#
30	Cattell Continuous Reaction Test - Total Score#
31	Vital Capacity#
33	Response to Breathing Resistance##
34	[4]Breathing Pattern##
35	Electroencephalogram##

* A positive standard score = superior performance
** A low crude score = positive standard score
\# A high crude score = positive standard score
Crude scores were not converted into standard scores on these tests. They are ratings.

---

[4]A complete description of the respiratory measures and breathing pattern and a detailed statistical analysis of these measures may be found in the following: Franzen, Raymond and Blaine, Louisa. Evaluation of respiratory measures for use in pilot selection. Washington, D.C.: Civil Aeronautics Administration Division of Research, Report No. 25, January 1944.

SECTION I

STATISTICAL TREATMENT OF THE PART I DATA

In this section of the report are presented those statistical analyses conducted with the population included in Part I of the Pensacola studies. The scores of Naval Air Cadets (varying in number between 336 and 373) on 12 psychological and psychomotor tests and on 21 physiological measures are the basis for these analyses.

In order to facilitate comparisons of the tests, the original scores of each test (except response to breathing resistance, breathing pattern, and [illegible]) were expressed in terms of the mean and sigma of the test: $X-M_x/\sigma_x$ and all analyses have been made using the resultant standard scores.[5] The three tests where standard scores were not used were originally scored by a four-place rating, not a continuous variable, hence the original scores were used in considering these variables.

In carrying out the statistical treatment of these data, the following analyses were undertaken:

1. An investigation of reliability (using chi-squared) of the sample by comparison of "arbitrary" halves of the population on each of the tests. This is to be distinguished from the reliability of the measurements which were not available. All cases used were ranked in serial order. One half is composed of even ranks. The other half is composed of odd ranks.

2. Intercorrelation (r and eta) of the nine psychological and psychomotor tests.

3. The significance of the difference between the distribution of the scores of selected groups of pilots and the distribution of all pilots (using chi-squared). The selected groups were washouts for aptitude, pilots who had had board appearances, and washouts for other than aptitude and a chance selection.

4. Consideration of selected tests for distinguishing washouts to determine the most efficient cut-off point, i.e., maximum elimination of poor cadets with minimum elimination of successful cadets (using chi).

5. Consideration of various combinations of the selected tests to determine a pattern (battery) of cut-off points with greater efficiency than that provided by any one test alone (using multiple chi).

THE RELIABILITY OF THE SAMPLE

The test scores may be examined for homogeneity of the sample and for consistency of administration by a comparison of the distribution of scores

---

[5] The standard scores on all subjects for each test are on file with the National Research Council Committee on Selection and Training of Aircraft Pilots.

for "arbitrary" halves[6] of the population,[7] i.e., in the absence of another similar universe of data to which the same measures may be applied, the parent population already measured is divided in half and the tests for significant differences carried out.

The Chi-squared formula, $\frac{1}{N_1 N_2} \sum \frac{(F_1 N_2 - F_2 N_1)^2}{F_1 + F_2}$, may be employed to measure the significance of a divergence between two samples drawn from the same universe. The probability (P-value) obtained from such a chi-squared is that of having by chance a difference as great as or greater than the existing (observed) difference, if the two such samples are withdrawn from a homogeneous parent population at random.

The chi-squared and P-value for each test appear in Table 1. For example, Tests 15 (Ophthalmograph - fixations per line) and 31 (Vital Capacity) represent the two extremes of difference and likeness between their halves. The latter has a P-value of .95 which means that by chance there would be as much or more difference between samples withdrawn from the same parent universe 95 times out of every 100 such comparisons. Test 15 with a P-value of .02 has such divergence in the distributions of its halves that only twice out of 100 comparisons of samples with the same origin would such a difference be due to chance alone. Test 31 may be considered to measure a homogeneous group with uniform standards of measurement. Test 15 must be mistrusted either as to its administration or as to the present sampling of the quality measured.

A P-value of .05 may be accepted as the point above which a difference is not considered significant and below which a difference cannot be ignored. The adequacy of the samples of Tests 5, 9, 14, and 15 must be questioned. The Two-Hand Coordination Test (Test 4) cannot be eliminated but would require repeated applications with greater consistency in results before the interpretations based on this application may be accepted. Consistency in this test is dependent upon calibration of the two-hand mechanism to uniform standards and this may be the cause of the unreliability in sampling.

All the other tests have P-values large enough to indicate that differences between their halves are probably due to chance. The test samples may therefore be regarded as homogeneous. The importance of this result must not be overlooked since a distinction in distribution between washouts and those who have been accepted cannot have meaning unless the material is sufficiently homogeneous to make chance "pockets" improbable. Use of chi-squared (or any other such procedure) to distinguish a selected group from its parent population must assume a homogeneous parent.

---

[6] Even ranks vs. odd ranks when arranged in serial order.

[7] Distributions in standard deviation intervals of the halves and of the total population of each test are on file with the Committee on Selection and Training of Aircraft Pilots.

DISTRIBUTION CHARACTERISTICS OF [illegible] DATA

Test Number	Comparison of Halves P	X^2	Degrees of Freedom	Total Distribution Parameters Mean	Sigma	N
1	.22	5.47	12	70.0	8.7	373
2	.88	6.70	12	52.2	8.7	336
3	.41	12.47	12	449.0	113.7	371
4	.06	19.69	12	59.3	11.9	371
5	.04	20.80	11	23.1	9.4	373
6	.73	8.73	12	5.97	1.03	374
7	.43	12.16	12	214.6	25.2	373
8	.79	7.91	12	43.1	8.4	339
9	.02	23.81	12	46.1	21.2	357
10	.66	9.52	12	70.0	8.2	372
11	.92	5.91	12	115.5	8.7	373
12	.41	12.52	12	66.9	8.2	373
13	.65	6.90	9	10.8	5.1	372
14	.03	20.92	11	370.5	114.9	368
15	.02	24.19	12	8.5	1.4	356
16	.48	10.61	11	1.6	14.5	367
17	.37	12.98	12	21.8	11.8	359
18	.17	16.62	12	17.6	10.1	359
19a	.19	10.02	12	28.4	9.3	361
19b	.22	13.10	10	37.5	10.7	361
19c	.51	10.21	11	17.9	7.9	361
19d	.35	12.16	11	12.1	5.7	361
24	.70	4.67	7	1.15	2.26	337
25	.95	2.65	8	5.54	6.37	337
26	.68	9.29	12	8.73	7.77	337
27	.24	15.00	12	424.0	80.1	340
28	.88	6.70	12	321.6	79.2	338
29	.93	5.78	12	2776.0	319.0	360
30	.94	6.38	12	746.0	144.0	336
31	.95	6.12	12	5262.0	689.0	363

*It should be noted that the X^2 for halves of the distribution will tend to be lower than that of the total distribution. This may be expected from the reduced size of the samples involved in the calculations.

Reliability correlations for Basal Air corrected for Body Surface (first two minutes with second two minutes, for instance) have been found to be over .90. The measurement is reliable, but the sample is not. Changing conditions of measurement may increase reliability coefficients, but lower the reliability of the sample.

INTERCORRELATION OF NINE PSYCHOLOGICAL AND PSYCHOMOTOR TESTS[8]

Since the psychological and psychomotor tests used were chosen because of their apparent pertinence in measuring a given set of coordination and reaction pattern thought to be basic to successful flight performance, it is necessary to investigate the extent to which they do measure the same or related qualities, i.e., the degree of correlation among their several results.[9]

Linear or curvilinear correlation (see Table 2) among the psychomotor tests is always less than .4. The highest relationship between any two tests is that between Tests 2 and 8, the Otis and Minnesota Paper Form Board, with an r of .36 and an eta of .39. The Mashburn Serial Action Scores (Test 6) are most apt to be associated with the scores on the other tests. The etas of this test are close to .3 with every other test, except Tests 5 and 9, which have been found to have doubtful sampling consistency. This is the only case of a relationship pattern in the correlation matrix.

If subsequent testing were to yield higher coefficients, the hypothesis of common factor, that is, the hypothesis that the Mashburn, for instance, evaluates certain aspects of all the qualities measured by the other six tests, would be worth investigation. Single order r's of .3, however, should not be used as a basis for multiple associations because any chance errors in relationship artificially increase the multiple coefficients. The present measures suggest the possibility that the Mashburn would be apt to have non-linear relations with the psychomotor tests but a linear relation with the Otis (as indicated by the size of eta minus r).

With the possible exception of Test 6 (Mashburn), it is apparent that these psychological and psychomotor tests, in so far as they are reliable, measure very different characteristics. It is also apparent that some of the relations tend to be non-linear. They must then be separately evaluated in terms of their function in pilot selection. They may be combined in such a manner as to obtain maximum predictive efficiency. The next parts of this section of the report will deal with this problem.

---

[8] The physiological measures were selected much more at random than the other tests in an effort to find physical characteristics that might have bearing on flying aptitude. Correlations among them, therefore, were treated in groups to be presented in subsequent reports. These groups are such as thirteen Tilt Table tests, nine Respiration measures, etc.

[9] The Cattell Tests were not included in these intercorrelations because they were not available until later in the study.

TABLE 2

INTERCORRELATIONS (r AND ETA) OF ALL PSYCHOLOGICAL AND PSYCHOMOTOR TESTS

Test No.		1	2	3	4	5	6	7	8	9
1	r	----								
	eta	----								
2	r	.141	----							
	eta	.224	----							
3	r	.171	-.010	----						
	eta	.251	.165	----						
4	r	.235	.245	.045	----					
	eta	.270	.338	.168	----					
5	r	.093	.138	.156	.041	----				
	eta	.201	.180	.323	.125	----				
6	r	.208	.290	.135	.297	.013	----			
	eta	.282	.297	.294	.343	.194	----			
7	r	.203	.054	.077	.170	-.090	.276	----		
	eta	.245	.236	.196	.263	.250	.333	----		
8	r	.108	.360	.068	.225	-.008	.253	.182	----	
	eta	.204	.385	.143	.274	.197	.332	.239	----	
9	r	.005	.127	.039	.155	.127	.043	-.116	-.005	----
	eta	.157	.217	.144	.233	.277	.207	.171	.203	----
No.		373	336	371	371	373	374	371	339	
Mean		70	52.2	449	59.3	23.1	5.97	215	43.1	
Sigma		8.7	8.7	113.7	11.9	9.4	1.03	25.2	8.4	

1. Eye-Hand Coordination Test
2. Otis Test of Mental Ability
3. Ataxiameter
4. Two-Hand Coordination Test
5. Perception of Change in Position
6. Washburn Serial Action Test
7. McDougall Dotting Test
8. Minnesota Paper Form Board Test
9. Thorndike-Kelley Athletic Achievement

COMPARISON OF SPECIFIC CRITERION GROUPS AND THE TOTAL POPULATION

Since the major objective of these investigations is to select tests which would fail those who lack aptitude for flying, each of the tests and measures must be evaluated in terms of the degree to which its scores distinguish those with certain unsuccessful flying experience from the total population of cadets tested.

By means of the chi-squared formula, difference between the distributions of test scores of certain selected samples and the total distribution of all cadets may be evaluated. The chi-squared formula, $\chi = \frac{(o-e)^2}{e}$, is used to express the amount of divergence between a sample and its parent population. The symbol "o" in the formula is the observed frequency of the sample and "e" i the theoretical frequency.[10] The probability (P-value) then states the chances that the group studied is a chance withdrawal from the parent population.

The distributions of test scores for three groups of cadets were compared with the parent population by means of this procedure:[11] (1) washouts for inaptitude, (2) those dropped from training for reasons other than aptitude, and (3) board appearances.[12]

The results of these comparisons are summarized in Table 3. This table presents the P-values of each of the groups studied for 31 of the psychological and physiological tests. In the last column of the table (P-value of chance selections) are the P-values for a comparable group of cadets selected at random from the total population and evaluated in the same manner as the various criterion groups.

The detailed data for each of the criterion groups are presented separately in Tables 4, 5, 6 and 7. When the P-values are below .03, the group described may be considered as differing significantly from the parent population with respect to the test analyzed. A difference of this degree would occur less than 3 times out of every 100 chance withdrawals from the parent group.

Great care must be exercised in the interpretation of these P-values. It will be observed that low P-values appear in the last column of the table in some cases. These are chance withdrawals from the total popula-

---

[10] This is the general formula. Special applications depend upon how "e" is determined. In this case the theoretical frequency is the "expected" frequency in the sense that the sample is expected to be distributed in the same proportions throughout the variable as is the parent population.

[11] Distributions on Sigma intervals of each criterion group and their theoretical (expected) distributions used in computing the chi-squared test of significance for each test are on file with the National Research Council-Committee on Selection and Training of Aircraft Pilots.

[12] A complete description of these criterion groups is presented in the final summary report of the Pensacola studies. See McFarland, Ross A. and Franzen, Raymond. Op cit.

tion being studied. If we withdrew 100 chance arrangements of the same size, then three of these are likely to show a P-value of .03 or less. For this reason it becomes essential to test the reliability of the P-values by computing it for halves in the manner indicated on page 4. Such an analysis will follow in this discussion. One of the low P-values (.016) for the chance withdrawals, it will be noted, is for Test 4, the Two-Hand Coordination Test (see Table 3). This again points to the possible lack of homogeneity in these figures as indicated earlier in the report by comparison of halves of the total distribution of test scores.

The first criterion group selected to represent flying inability was a group made up of cadets who had been "washed out" because of inaptitude for flying, demonstrated during the pilot training period. The chi-squared and P-values for this group, presented in Table 4, should be read as follows:

In the case of Test 10 (Diastolic Blood Pressure -- reclining) the distribution of the scores of 35 washouts was so similar to the distribution of the scores of all pilots that differences at least as large as those which occurred could be expected 97 times out of 100 in samples withdrawn by chance from the same universe of data. (The P-value for a chi-squared of 1.61 with 7 degrees of freedom is .97.) It is obvious that failure in flight training is not significantly associated with performance on this measure. Test 6 (the Mashburn Serial Action Test), on the other hand, yields scores for washouts that are distributed in a manner so different from the scores of all pilots that only three times out of 10,000 would chance produce a difference as large or larger in withdrawals from the same universe of data. (The P-value for a chi-squared of 23.52 with 6 degrees of freedom is .00028). It may be assumed therefore that Test 6 does distinguish the washouts from the total populations, i.e., the test behavior on the Mashburn Serial Action Test of those cadets eliminated from flight training is of a significantly different nature than is the behavior of all pilots.

Assuming the P-values to be reliable, Tests 2, 4, and 6 (the Otis, Two-Hand Coordination, and Mashburn), and also possibly 13 and 19d (the Schneider Index and Time Interval to smallest pulse pressure on the Tilt Table Test), would be selected as measures which would significantly differentiate failures in flight training (for reasons of inaptitude) from the total population of cadets.

In the comparison of halves, the reliability of the Two-Hand Coordination Test (Test 4) was not satisfactorily demonstrated (see page 5). The probability of chance causing the difference between washouts and total population is only slightly lower than the P-value of the halves, and might, therefore, be due to the known lack of consistency in the test's administration. There is, however, good reason, because of the nature of the test, not to eliminate it as a possibility for pilot selection, but rather to enforce better calibration and standard procedures of administration. This was done at a later date. All subsequent considerations of the Two-Hand Coordination Test are on the assumption that standard conditions of measurement would provide

results with adequate reliability.[13]

Inspection of the observed and expected frequencies[14] indicated the nature of the large difference between washouts and total pilots in their Mashburn scores. Superior performance (scores above the mean) would have been expected among approximately 19 of the 35 washouts. Actually only 8 of them scored above average. Twice as many washouts had scores that were .7 or more standard deviations below the mean as would have been the case had their distribution resembled that of all pilots. These differences appeared at all intervals of the distribution.

In the case of the Otis Intelligence Test (Test 2), however, the differences tend to be at both extremes but not in the middle of the distributions. Nine washouts, as compared with 4.4 expected cases, had extremely low scores, i.e., 1.8 or more standard deviations below the mean. Only 3, instead of the 9.1 expected frequency, occur at .7 or more standard deviations above average. The other 22 cases are distributed according to expectation.

The character of these relative distributions -- the psychomotor tests scoring washouts lower throughout the distribution and the intelligence test distinguishing them at the two extremes -- suggests the possibility of compensation among the tests. It may be that very high intelligence will allow for pilot maintenance, even if other abilities are absent and, conversely, very low intelligence (low in terms of the group tested) will outweigh other abilities and make for poor flying aptitude. Psychomotor reactions (measured by the Mashburn Test) may be significantly associated with pilot inaptitude whenever they are below average.

There are two other groups of cadets whose flight proficiency had been under question but who were not poor enough to be failed for reasons of inaptitude. One is composed of cadets who were dropped from training on grounds described as "other than lack of flying aptitude," and the other is composed of cadets who had "board appearances" but who had not, as a result, been grounded. Comparison of selected scores of these groups of cadets and the total population are presented in Tables 5 and 6. The tests tentatively selected on the basis of the P-values for washouts

---

[13] Editor's Note. Subsequent to the writing of the report, this test was revised. For a discussion of the original and revised forms of the Two-Hand Coordination Test and of data obtained with this equipment, see: McFarland, Ross A. and Channell, Ralph C. A revised two-hand coordination test. Washington D. C.: Civil Aeronautics Administration Airman Development Division, Report No. 36, October 1944.

[14] These distributions are on file with the National Research Council Committee on Selection and Training of Aircraft Pilots.

TABLE 3

SUMMARY OF P-VALUES FOR CRITERION GROUPS AND CHANCE SELECTIONS WHEN COMPARED WITH ALL PILOTS

Test	P of Washouts for Inaptitude	P of Cadets Dropped for Reasons Other than Aptitude	P of Board Appearances	P of Chance Selections
1. Eye-Hand	.166	.201	.155	.154
2. Otis	.030	.865	.659	.728
3. Ataxiameter	.804	.711	.860	.267
4. Two-Hand	.043	1.000	.260	.016
5. Change in Position	.808	1.000	.548	.930
6. Mashburn	.00028	.726	.135	.307
7. Dotting	.590	.889	.062	.440
8. Paper Form Board	.436	.246	.156	.391
9. Athletic Achievement	.202	.779	.676	.544
10. Diastolic B. P.	.973	.779	.0090	.939
11. Systolic B. P.	.406	.575	.534	.686
12. Pulse Rate	.700	.317	.933	.986+
13. Schneider	.048	.067	.00049	.011
14. TA/BS	.914	.159	.553	.653
15. Ophthalmograph	.328	.174	.160	.275
16. B.M.R.	.334	.596	.859	.186
17. Cold Pressor, diast.	.759	.004	.373	.910+
18. Cold Pressor, systolic	.619	.765	.188	.783
19a Tilt P.P. Change	.683	.093	.391	.036
19b Tilt P.R. Change	.782	.509	.725	.680
19c Tilt Smallest P.P.	.252	.435	.082	.064
19d Tilt Time Score	.056	.242	.788	.429
24. Startle-Duration	.607+	.343	.607+	.607+
25. Startle-Amplitude	.505	.536	.842	.888
26. Startle-Dur. Change	.392	.076	.065	.200
27. Cattell I	.472	.253	.183	.191
28. Cattell II	.425	.505	.279	.961
29 TC/BS	.377	.805	.271	.351
30. Cattell Total	.309	.334	.279	.417
31. F.I	.3[illegible]	.[illegible]	.826	.539

Test Order	P	χ^2	Degrees of Freedom	Test Scores for Aptitude	N of Percent Distribution
1	.166	6.52	4	35	373
2	.030	12.41	5	34	336
3	.804	3.04	6	35	373
4	.043	11.40	5	35	373
5	.808	3.70	6	35	373
6	.00028	23.52	5	35	374
7	.590	3.73	5	35	371
8	.436	5.89	6	34	339
9	.202	9.79	7	35	357
10	.973	1.61	7	35	372
11	.406	4.00	4	33	373
12	.700	3.00	5	35	373
13	.048	11.18	5	35	372
14	.914	2.05	6	35	360
15	.328	5.80	5	30	356
16	.834	2.77	6	34	367
17	.759	3.37	6	31	359
18	.619	3.54	5	31	359
19a	.683	3.95	6	33	361
19b	.782	2.45	5	33	361
19c	.252	5.40	4	33	361
19d	.056	12.32	6	33	361
24	.607+	.55	2	33	337
25	.505	4.33	5	33	337
26	.892	2.90	7	33	337
27	.472	4.58	5	30	340
28	.423	4.95	5	30	338
29	.077	8.48	4	33	360
30	.309	4.82	4	29	336
31	.394	4.10	4	34	363

"N's appear in Tables 4, 5, 6, and 7. N's for all groups are approximately 35 except the "dropped-for-reasons-other-than-aptitude" group which is approximately 12.

TABLE 3

COMPARISON OF CADETS DROPPED FOR "REASONS OTHER THAN APTITUDE" AND ALL PILOTS ON EACH TEST

Test Number	P	Chi*	N of Dropped for Other than Aptitude	N of Parent Distribution
1	.241	1.28	13	373
2	.815	.17	12	336
3	.711	.37	13	373
4	1.000	0	12	371
5	1.000	0	13	373
6	.736	.35	13	374
7	.839	.4	13	371
8	.246	1.16	12	339
9	.779	.28	13	357
10	.779	.28	13	372
11	.575	.56	13	373
12	.317	1.00	13	373
13	.067	1.83	12	372
14	.159	1.41	12	368
15	.174	1.36	12	356
16	.596	.53	12	367
17	.004	2.89	13	359
18	.765	.30	12	359
19a	.073	1.79	13	361
19b	.509	.66	13	361
19c	.435	.78	13	361
19d	.242	1.17	13	361
24	.343	.95	12	337
25	.536	.62	12	337
26	.826	.22	12	337
27	.353	.93	12	340
28	.523	.64	12	338
29	.889	.14	12	360
30	.184	1.33	12	336
31	.569	.57	13	363

*Chi, not chi-squared, was employed in these calculations because the small N of the selected group does not allow more than 2 intervals in the distribution and, therefore, only one degree of freedom.

(i.e., Tests 2, 4, and 6) make absolutely no distinction of the "dropped" pilots and reveal no significant divergence by board appearances. It is interesting, however, that the Washburn Test (Test 6) which has seemed to hold the greatest possibilities has a relatively, though not significantly low P-value in the case of board appearances.

It will be noted that those psychological and psychomotor tests which furnished important clues to the selection of washouts provide no description of those pilots selected on grounds other than decided inaptitude for flying. The only suggestions that the chi-squared measures offer are among the physiological tests where three cardiovascular measures (Tests 10 and 13 among board appearances and Test 17 among dropped pilots) have a very low probability that the difference between the selected and total group is due to chance. It is worthy of note that the only measures with a P-value under .1, in the case of "dropped" pilots, are Tests 13, 17, and 19a.

The problem, however, is confused by the unreliability of the physiological tests as measures of individuals, though the scores may be perfectly reliable representations of physiological moments in the individual's experience. It is because these tests are measuring processes which vary with conditions, that unreliability must exist until sufficient repetition establishes norms for individuals that truly distinguish them in terms of physiological function. One application of such tests measures only the physiological instant, which is determined by innumerable other uncontrollable influences in the organism and in the environment.

In spite of this limitation, the three tests with distinctive P-values (Diastolic Blood Pressure -- reclining; Cold Pressor -- greatest diastolic change, and the Schneider Index), measure physiological functions which might be associated with emotional excitation. It seems possible, therefore, to regard them as suggestive of behavior that will call a pilot into question even though he is considered technically competent. The two physiological tests (13 and 19d) that seemed to indicate some possibility of distinguishing washouts were also cardiovascular measures: the Schneider Index and the Tilt Table -- time interval to smallest pulse pressure.

It is very important for future research on physiological measures that a reliable and valid index of cardiovascular efficiency be developed. These materials definitely indicate that such a test could play an important role in the selection of pilots. All conclusions regarding the selective efficiency of these tests must, however, be drawn with extreme caution. Extensive cross validation of these results and standardization of the physiological measures themselves must be undertaken before results of their use will be trustworthy.

If any given test, purported to distinguish inaptitude for flying because its P-value indicated such a distinction, could occur by chance only twice out of 100 such comparisons, then, since only one instance is represented by these calculations, it is always possible that the comparison in question is one of those two times. The less homogeneous the population, the more is any given divergence of sample from the parent population apt to be a chance one, even though the probability of such a chance occurrence is computed as low.

TABLE 6

COMPARISONS OF CADETS WITH "BOARD APPEARANCES" AND ALL PILOTS ON EACH TEST

Test Number	P	X^2	Degrees of Freedom	N of Board Appearances	N of Parent Distribution
1	.155	8.02	5	36	373
2	.659	2.43	4	31	336
3	.860	2.54	6	36	371
4	.260	7.74	6	35	371
5	.548	4.97	6	36	373
6	.135	9.79	6	36	374
7	.062	10.57	5	36	371
8	.156	8.01	5	31	339
9	.676	3.16	5	29	357
10	.0098	15.15	5	35	372
11	.534	3.10	5	36	373
12	.933	1.26	5	36	373
13	.00049	22.17	5	34	372
14	.553	3.03	4	34	368
15	.160	7.94	5	33	356
16	.859	2.55	6	35	367
17	.373	5.39	5	32	359
18	.188	6.18	4	32	359
19a	.391	5.23	5	36	361
19b	.725	2.83	5	36	361
19c	.082	11.24	6	36	361
19d	.788	3.16	6	36	361
24	.607	.55	2	30	337
25	.842	1.39	4	30	337
26	.065	8.87	4	29	337
27	.183	8.86	6	34	340
28	.279	5.09	4	34	338
29	.271	6.41	5	30	360
30	.270	6.42	5	33	336
31	.926	1.90	6	31	363

This limitation applies to the data under consideration in these analyses. It is important, therefore, to test the reliability of any P-value. One reflection on the degree of this limitation is to compare halves of the population as reported earlier in this report. Another reflection, empirical in nature, is to select at random 36 pilots (about the number of washouts) and applying the same chi-squared technique to the distributions of their scores as was employed in the analysis of the groups with specific flying experience. These results are merely by way of illustration and can not be taken as proof. The results of this analysis are presented in Table 7.

Test 4 (the Two-Hand Coordination Test), which, because of its testing technique, was found to have doubtful reliability in the comparison of its halves, is found to have a larger difference (smaller P-value) between chance withdrawals and the total distribution than between the distributions of washouts and total. It may be tentatively concluded that this test, with better calibration and standardized conditions of application, would have predictive value, but that as administered at the present time samplings may be uncomparable.

Similarly, Test 13 (Schneider Index) distinguishes between chance selection and the parent population (P = .01) more significantly than between washouts and the parent population (P = .05), but less truly than between board appearances and the parent population (P = .0005). On the other hand, comparison of halves in this test has a P-value of .65.

The remaining tests that offered possibilities, 2, 6, 10, 17 and 19d, (Otis, Mashburn, Diastolic Blood Pressure -- reclining, Cold Pressor -- greatest diastolic change, and Tilt Table Test -- time interval to smallest pulse pressure) all show satisfactory homogeneity by both methods.

This comparison of chance selections with the total population merely illustrates the difficulty. If an extreme difference occurs then there is reason for suspicion. If the chance selection shows a P-value indicating a large probability that such a sample will appear by chance, this does not constitute any favorable evaluation. The comparison of halves does test homogeneity. The evaluation of chance selections may illustrate heterogeneity.

Analyses of Tests 33, 34, and 35 (Response to Breathing Resistance, Breathing Pattern, and EEG) receive only limited attention in this report. The differences between both the washouts and board appearances and the parent population are not statistically significant. The reason for separate and limited treatment of these results is that the purpose of this study is evaluation of the selective function of tests. Ratings are, by their very nature, inappropriate for selection. In the first place, ratings are not objective and in the second place, they (4-place ratings like these) select too many cases in their lowest scale score to be practical.[15]

---

[15] A complete analysis of breathing records was presented in an earlier report. See Franzen, Raymond and Blaine, Louisa. Op. cit. A detailed analysis of electroencephalographic records will be made the subject of a subsequent report. One such analysis has already appeared in this series of technical reports. See Forbes, Alexander and Davis, Hallowell. Electroencephalography of naval aviators. Washington, D. C.: Civil Aeronautics Administration Division of Research, Report No. 13, April 1943.

TABLE 7

THE PROBABILITY THAT PILOTS WHO HAVE BEEN SELECTED BY CHANCE ARE NOT SIGNIFICANTLY DIFFERENT FROM ALL PILOTS IN THEIR SCORES ON EACH TEST

Test Number	P	X^2	Degrees of Freedom	N of Chance Selection	N of Parent Distribution
1	.154	9.41	6	36	373
2	.728	3.61	6	34	336
3	.267	6.46	5	36	371
4	.016	13.99	5	36	371
5	.930	2.40	7	36	373
6	.307	5.99	5	36	374
7	.440	5.86	6	36	371
8	.391	6.31	6	34	339
9	.544	5.00	6	34	357
10	.939	1.70	6	36	372
11	.686	3.93	6	36	373
12	.986+	.89	6	36	373
13	.011	16.60	6	36	372
14	.653	4.18	5	36	368
15	.275	6.37	5	33	356
16	.186	6.20	4	36	367
17	.910+	.36	4	35	359
18	.783	1.73	4	35	359
19a	.036	11.92	5	35	361
19b	.880	1.73	5	35	361
19c	.064	10.49	5	35	361
19d	.429	5.95	6	35	361
24	.607+	.15	2	33	337
25	.888	1.66	5	34	337
26	.200	5.99	4	34	337
27	.191	8.73	6	34	340
28	.961	1.38	6	31	338
29	.351	6.71	6	34	360
30	.417	6.06	6	31	336
31	.539	4.08	5	34	363

CONSIDERATION OF SELECTED TESTS TO DETERMINE STANDARD OF ELIMINATION FOR INAPTITUDE

Thus far in this report the difference between washouts and all cadets has been considered as it exists throughout the entire range of the distribution of test scores. The problem up to this point has been to find tests that measured some characteristics in which the unsuccessful fliers were different from fliers as a whole. Ideally, however, the problem is to divide the population into a dichotomy of flying success and flying failure. It is not one of considering varying degrees of flying ability as related to a test's score, but of locating a level in any test or combination of tests that will distinguish between those cadets who will in all probability be eliminated from flight training and those who will succeed.

Practically, this cut-off point cannot be located so as to eliminate all potential failures and yet retain all potential successes, but must be put at the level of greatest efficiency, i.e., rejecting the maximum number of the "failure" category and the minimum number of the "success" category.

The chi technique,[16] applied to the washouts vs. total population group, when cumulated into pass and fail categories at any score level, will test for greatest selective efficiency by showing the cut-off at which the pass and fail division of unsuccessful pilots is most different from the same division of all pilots. Such levels must then be tested for the proportion of successful and unsuccessful pilots that they would eliminate. Thus a given cut-off might yield a very high chi because it "failed" 80 per cent of the washouts and 40 per cent of all pilots. But it is not a satisfactory level since 40 per cent failure of all pilots are neither efficient nor desirable.

In view of the recognized opportunity for a chance determination of such chis, the materials were again divided in the manner described on pages 3 and 4, i.e., a chi was computed separately for halves of each test. Unless the chi at any level tested is high for both halves, that level should not be accepted. These chis are presented in Table 8. It is not necessary to express these chis in terms of probability. The chi may be interpreted just as a standard score (X/σ_X) value would be. A negative chi means that there is a smaller proportion of washouts than of the parent population below the elimination level.

A chi of 1.0 indicates a probability of .3 that the difference was the result of chance; a chi of approximately 1.90 represents a P-value of .05 (.025 at each end of the distribution), etc. Therefore, with some latitude, the chis for both halves of any test level must be as large as this or larger before the test meets an acceptable criterion of distinguishing between washouts and total population by more than chance.

---

[16] Chi, not chi-squared, since the categorical grouping reduces the degrees of freedom to one and probability, then, is the same for chi as for X/σ_X. See Franzen, Raymond. A method for selecting combinations of tests and determining their best "cut-off points" to yield a dichotomy most like a categorical criterion. Washington, D. C.: Civil Aeronautics Administration Division of Research, Report No. 12, March 1943.

Unfortunately, if this requirement is rigidly enforced only two tests cut off at the mean would meet it. Recognizing that errors of measurement will lower chi values, particularly when the size of the sample has been reduced (as is done when halves are compared), the three psychological tests 2, 4, and 6 (Otis, Two-Hand, and Mashburn) at standard score levels of -.7, -.4, and -.4 respectively, were accepted as reliable and significant distinctions between washouts and total. All chi values for halves are well over 1. Here, again, is apparent the different discriminative function of the Otis Intelligence Test (Test 2) and the psychomotor tests. It is the lower level of intelligence that is more characteristic of washouts than of successful pilots. With the Mashburn and Two-Hand, it is necessary to set the failure level nearer the mean to eliminate a significantly different proportion of washouts and successes. Both the reliability and significance of the psychomotor test are increased if the break between success and failure is moved up to the mean.

Test 4, the Two-Hand Coordination Test, which demonstrated such low reliability because of known limitations in its administrative mechanism, shows up very well in the relative values of the chis of halves. Since the source of its lack of homogeneity is known and since the abilities tested may easily be associated with the manipulation of a ship's controls, it is retained in the subsequent discussion of the best selective tests. This is done, however, with full realization that no use may be made of the test for selection purposes until consistencies in its application can be demonstrated.

The three psychological tests are the only measures which exhibit chis sufficiently large, and in the same direction in both halves, to distinguish reliably the washouts from all pilots at any success and failure (standard score) level.

TABLE 8

CHIS* FOR WASHOUTS AGAINST ALL PILOTS COMPUTED FOR TWO HALVES (A & B) OF EACH TEST

Test No.	Halves	Chis for two-place comparison when cut-off is at and including: -1.8 ss	-1.3 ss	-1.0 ss	-.7 ss	-.4 ss	-.1 ss
1	A				.56	.90	1.07
1	B	1.71	1.86	1.35	.94	1.41	.93
2	A		1.81	1.11	2.33	2.74	2.33
2	B		1.40	.78	1.14	.96	1.42
3	A			.42	-.39	.35	-.44
3	B			.66	.24	.36	.69

* Chi, not chi-squared, because of only one degree of freedom.

TABLE 8 (Cont'd)

Test No.	Halves	-1.8 ss	-1.3 ss	-1.0 ss	-.7 ss	-.4 ss	-.1
4	A			1.50	.42	1.66	2.
4	B		.75	1.82	2.40	2.66	2.
5	A**			0	-.71	-.14	1.
5	B**				-.82	0	.1
6	A			-.20	1.14	1.22	2.
6	B		2.06	1.82	2.29	3.96	3.0
7	A		0	-.74	-1.36	-.33	-.
7	B		.66	1.19	1.01	1.36	2.
8	A	1.82	1.73	2.09	3.02	1.80	1.
8	B				-.90	.41	.
9	A	.75	.57	-.24	-.37	-1.37	-.5
9	B	2.56	2.39	2.79	1.92	1.52	1.8
10	A			-.10	-.14	1.20	.
10	B		1.17	.63	-.14	0	.
11	A		.53	.41	-.41	.36	.4
11	B		1.42	1.51	.77	1.95	2.81
12	A		1.41	.35	.20	1.13	.66
12	B		-.22	.14	-.28	.51	.4
13	A				.17	-.62	.10
13	B				1.30	1.51	.8
14	A				-1.04	.33	.4
14	B		2.52	1.53	1.64	.49	.4
15	A		.54	-.14	.57	1.39	.9
15	B				-1.20	-1.32	-1.06
16	A			1.40	.70	.77	.77
16	B			.24	.14	.46	.20
17	A			.14	0	.24	.68
17	B				0	.45	.62
18	A				-.30	-1.03	.14
18	B		1.42	.42	.50	-.10	-.36

**Reverse signs of standard scores at cut-off points.

TABLE 8 (Cont'd)

Test No.	Halves	-1.8 SS	-1.3 SS	-1.0 SS	-.7 SS	-.4 SS	-.1 SS
19a	A			.22	1.19	1.30	.58
19a	B			.50	.28	.57	1.91
19b	A		2.00	1.54	1.45	.96	.20
19b	B			.30	0	-.78	-.14
19c	A		1.32	.78	1.42	1.50	.92
19c	B		1.42	1.54	1.44	1.14	.14
19d	A**			1.34	1.64	.84	.53
19d	B**			.77	.14	.17	-.14
24	A					.14	.35
24	B	2.08	1.42	1.30	2.17	1.95	1.30
25	A					-.47	-.58
25	B		.95	.76	1.76	.78	1.61
26	A					-.52	-.14
26	B		.44	2.32	1.38	1.48	1.[illegible]
27	A	2.82	3.70	2.79	2.51	1.57	1.37
27	B			-.28	-.68	1.09	1.30
28	A	3.14	3.31	1.65	1.63	1.61	1.51
28	B		0	-.34	.01	-.01	-.28
29	A			1.51	1.12	.96	2.45
29	B			.48	.10	.78	-.20
30	A	3.17	4.01	3.12	2.73	1.59	1.35
30	B		-.82	-1.14	-1.25	0	1.15
31	A	.72	.94	1.19	1.37	.70	1.17
31	B		.64	.49	.33	-.64	-1.02

**Reverse signs of standard scores at cut-off points.

The efficiency of the cut-off levels in terms of elimination of maximum washouts and minimum non-washouts, is apparent in the comparison of per cent of washout and remaining cadets "failed" at each interval of the standard deviation of Tests 2, 4, and 6, given in Table 9. The same percentage for Tests 1, 19c, 19d, and 29 (those next in recommendation from previous analysis) are also given so as to point out the greater selective efficiency of the three chosen tests.

As a rough standard of efficiency, assume for a moment that, in order to be considered, a test level must fail at least half of the potential failures (washouts) while failing no more than 20 per cent of those who would be retained by experience (non-washouts).

The categorical chis pointed to -.7 standard score on Test 2, -.4 on Test 4, and the same on Test 6, as the lowest levels with the greatest reliable difference between washouts and total. All three tests approximate the efficiency standard at these levels with eliminations of:

	Washouts %	Total %
Standard	50 or more	20 or less
Test 2	44	23
Test 4	60	33
Test 6	57	27

No one of the three tests meets the conditions of efficiency in proportions failed. Although they all show a significant difference between washouts and parent population in total distributions and in categorical divisions at more than one level, they all fail too many successful pilots at points where they will fail a sufficient number of those without flying aptitude.

It has been shown that each of the tests alone approximates a good selection device, but it is not known whether failure on all three, or on any two and not the other, may not provide a mechanism which will distinguish with adequate efficiency between men who will or will not be successful pilots. That is, up to this point the tests have been considered individually. It remains to be shown in the following portions of the report how the tests operate in various combinations at various score levels.

TABLE 9

PER CENT OF WASHOUTS AND OF REMAINING CADETS WHO WOULD BE ELIMINATED BY EACH STANDARD SCORE CUT-OFF ON SELECTED TESTS

	Test 1		Test 2		Test 4		Test 6		Test 19c		Test 19d*		Test 29	
	W.O.	R.C.	W.O.	R.C.	W.O.	R.C.	W.O.	R.C.	W.O.	R.C.	W.O.	R.C.	W.O.	R.C.
.8 & below	9	4	12	3	6	5	6	5	9	6	-	-	3	2
.3 & below	14	9	27	7	14	9	14	10	18	9	9	8	12	6
0 & below	20	15	27	17	29	13	26	17	27	16	33	8	21	12
- & below	34	25	44	23	37	21	40	21	36	21	42	37	27	19
& below	46	31	59	34	60	33	57	27	46	29	49	47	42	29
.1 & below	54	41	68	42	74	44	71	38	46	39	49	51	64	44
N	35	338	34	302	35	336	35	339	33	328	33	328	33	327

* Reverse signs of standard scores at cut-off points.

A COMPOSITE INDEX OF PASS AND FAIL

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 either test alone. If passing either of two tests compensates for failing the third, then pass-fail-fail, fail-pass-fail, and fail-fail-pass must be investigated. Failure on all three tests will be required for distinction between aptitude and inaptitude if the ability measured by one test compensates for the handicaps reflected in failure on the other two.

As part of this problem of the proper combination of passing and failing 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 must be considered at all possible success and failure cut-off points in order to find what combination of cut-off points will yield the largest number of failures among those rejected by experience accompanied by the smallest number of other failures.

This problem is one of testing the difference between the proportion of washouts who pass or fail and the proportion of non-washouts who pass or fail at various definitions of failure which include all three tests. These definitions may be at different levels of the distribution for each of the tests when they are combined.

How May We Find the Best Definition of Failure?

Table 8, presenting the two-place chis for each test at various cut-off points, shows a comparison of the washouts with the total group of pilots. The chis were obtained by the formula that compares a sample with its parent population, where the distribution of the sample (of the criterion group) was compared with the distribution that might have been expected from the distribution of the population of which it was a part. By this means, the value of each individual test for pointing out certain known characteristics of the samples may be demonstrated.

The problem here is the comparison of two groups of individuals purported to be different in kind. That is, the analysis concerns the isolation of a set of criteria that will eliminate those cadets with known inaptitudes without eliminating those who are not so handicapped.

Therefore, the chi-squared formula for comparing two independent groups (used in measuring the difference between the halves of each test) is employed as it is applied to a two-place comparison.[17]

---

[17] A complete discussion of the theory and application of the "Multiple Chi" technique is presented in an earlier report in this series. See Franzen, Raymond. Op. cit. Much of the material presented in Report No. 12 is reproduced in these pages to aid the reader in following the discussions.

Let the four values in the two categories be indicated thus:

	Failed by test	Not Failed by test	
Washouts	a	b	(a & b)
Remaining Cadets	c	d	(c & d)
	(a & c)	(b & d)	(a + b + c + d)

In this situation the chi formula for two independent groups 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	
Washouts	3	31	34
Remaining Cadets	42	255	297
	45	286	331

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

Table 10 on the following page gives pass and fail categories and chis obtained for all possible combinations of the three selected tests at various 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 washouts failed by test is less than the expected value.

It has already been shown that the same level of failure on all three tests may not be most efficient. Chis were computed for all seven combinations when the failure level was -.1 on two of the tests, that is, each of three combinations of two, but was -1.0 on the other, for all the seven combinations with -.1 failure level on each half 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. Other combinations and permutations are possible but these 63 should provide sufficient evidence.

The chis presented in the preceding table are in reality partial chis, for by combining the fail categories of different combinations we may obtain

TABLE 10

COMPARISON OF PASS AND FAIL CATEGORIES OF WASHOUTS AND OF REMAINING CADETS IN ALL COMBINATIONS OF THREE SELECTED TESTS AT VARIOUS LEVELS OF FAILURE

	Failure Level on All Three Tests:							
	-1.0 & less		-.7 & less		-.4 & less		-.1 & less	
Combination of Tests	Failed	Not Failed	Failed	Not Failed	Failed	Not Failed	Failed	Not Failed
(a) 2.46								
Washouts	2	32	4	30	3	31	2	
Remaining cadets	31	266	35	262	42	255	33	
Chi	-.84		-.003		-.86		-.94	
(b) 4.26								
Washouts	6	28	5	29	5	29	5	
Remaining cadets	19	278	29	268	33	264	32	
Chi	2.35		.90		.62		.69	
(c) 6.24								
Washouts	3	31	3	31	2	32	2	
Remaining cadets	29	268	30	267	23	274	26	
Chi	-.18		-.24		-.39		-.57	
(d) 24.6								
Washouts	2	32	3	31	2	32	3	
Remaining cadets	6	291	13	284	27	270	36	
Chi	1.39		1.15		-.62		-.57	
(e) 26.4								
Washouts	4	30	6	28	6	28	5	
Remaining cadets	10	287	15	282	17	280	21	
Chi	2.31		2.85		2.59		1.57	
(f) 46.2								
Washouts	2	32	3	31	4	30	4	
Remaining cadets	11	286	13	284	19	278	29	
Chi	.62		1.15		1.17		.37	
(g) 246								
Washouts	-	34	2	32	8	26	13	
Remaining cadets	4	293	7	290	20	277	38	2
Chi	-.68		1.20		3.33		3.89	

2.46 means failure is defined as below failure level on the Otis but above failure level on the Two-Hand Coordination and on the Mashburn Test.

24.6 means failure is defined as below failure level on the Otis and Two-Hand but above failure level on the Mashburn Test.

any of the following composite determinations of failure at any cut-off point. Call:

2.46	(failure on 2 but pass on 4 & 6)	a
4.26	(failure on 4 but pass on 2 & 6)	b
6.24	(failure on 6 but pass on 2 & 4)	c
24.6	(failure on 2 & 4 but pass on 6)	d
26.4	(failure on 2 & 6 but pass on 4)	e
46.2	(failure on 4 & 6 but pass on 2)	f
246	(failure on 2 & 4 & 6)	g

We may obtain a chi for:	By combining the failures of:
Failure on 2 without consideration of 4 & 6	a,d,e, & g
Failure on 4 without consideration of 2 & 6	b,d,f, & g
Failure on 6 without consideration of 2 & 4	c,e,f, & g
Failure on at least one of 2 or 4 without consideration of 6	a,b,d,e,f, & g
Failure on both 2 & 4 without consideration of 6	d & g
Failure on at least one of 2 or 6 without consideration of 4	a,c,d,e,f, & g
Failure on both 2 & 6 without consideration of 4	e & g
Failure on at least one of 4 or 6 without consideration of 2	b,c,d,e,f, & g
Failure on both 4 & 6 without consideration of 2	f & g
Failure on at least one of 2 or 4 or 6	a,b,c,d,e,f, & g
Failure on all three of 2 & 4 & 6	g

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
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[18]
Rejected at 2,4,6, & 7	25	9	34	3.10
Retained at 2,4,6, & 7	135	162	297	

[18] These two chis and the ones following may be compared with the cumulative two-place chis for -.1 cut-off given in the table "Chis for Washouts against all Cadets at a series of Cut-offs." They are slightly different because the formula used there was one comparing a sample with its parent population.

Similarly, all the other chis for a failure level of -.1 standard deviation are:

	Chi
6 without consideration of 2 or 4	3.61
At least one of 2 or 4 without consideration of 6	3.57
Both 2 & 4 without consideration of 6	2.75
At least one of 2 or 6 without consideration of 4	2.73
Both 2 & 6 without consideration of 4	4.32
At least one of 4 or 6 without consideration of 2	3.79
Both 4 & 6 without consideration of 2	3.48
At least one of 2 or 4 or 6	3.53
All three of 2 & 4 & 6 (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 combination 246, which is the same) than any of the partials 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, lower chis are obtained.

With the failure level set at the average, failure on all three tests or on both Otis and Washburn, while neglecting the Two-Hand Coordination Test makes the largest difference between proportion of washouts and non-washouts who are so failed.

These combinations may be tested for lower cut-off points in the interest of achieving more efficient proportions of failure.

At -.4 standard deviation:	Chi
2 & 6 without consideration of 4	4.39
2 & 4 & 6	3.33

At -.7 standard deviation:	
2 & 6 without consideration of 4	3.10
2 & 4 & 6	1.20

Thus it would not seem advisable to go further than half a standard deviation below the mean.

The question remains as to which of these composite failure criteria has adequate efficiency by our standard of eliminating 50 per cent of the washouts with only 20 per cent of the non-washouts. These composites are presented in Table 11.

TABLE 11

COMPARISON OF THE EFFICIENCY OF VARIOUS FAILURE LEVELS

Failure Levels	Chi	% of Washouts Failed	% of Remaining Cadets Failed
Below -.4 on 2 & 6 (without consideration of 4)	4.39	41	12
Below -.1 on 2 & 6 (without consideration of 4)	4.32	53	20
Below -.1 on 2 & 4 & 6	3.89	38	13
Below -.1 on either 4 or 6 (without consideration of 2)	3.79	94	61
Below -.1 on 6 (without consideration of 2 & 4)	3.61	71	38*
Below -.1 on either 2 or 4 (without consideration of 6)	3.57	94	64
Below -.1 on 2 or 4 or 6	3.53	100	72
Below -.1 on 4 & 6 (without consideration of 2)	3.48	50	23
Below -.4 on 2 & 4 & 6	3.33	24	7
Below -.1 on 4 (without consideration of 2 or 6)	3.10	74	45*
Below -.1 on 2 & 4 (without consideration of 6)	2.75	47	25
Below -.1 on either 2 or 6 (without consideration of 4)	2.73	85	62
Below -.1 on 2 (without consideration of 4 & 6)	2.72	68	43*

* The failure percentages for each test alone at -.1 differ from those in Table 12 because in the analysis of the tests in combination only those cadets who had taken all three tests would be included.

A sigma score of -0.1 on both Tests 2 and 6 (Otis and Washburn Tests) is the only failure level that so far has met the efficiency requirement of 50 per cent desired with 20 per cent undesirable failures. Nevertheless, policy may determine which of the composites to adopt. If the need is to fail as many as possible of the cases who would be washed out, even at the sacrifice of a large percentage of the successes or potential successes, then below average (-0.1) on Test 6 alone would probably prove adequate. If the problem is the more usual one of failing a maximum of the potentially successful cases together with a minimum of the potentially successful, then below -.1 or -.4 on both the Otis and Washburn, neglecting the Two-Hand, is the most acceptable standard. If, on the other extreme, a minimum sacrifice of cases is essential, then scores below -.4 standard deviation on all three tests should be required for failure.

The table following gives the "partial chis" when different levels of failure on the three tests are combined. So far, the analyses have considered the tests in all possible combinations when the failure level was the same on all three tests. It is possible that failing at -1.0 standard deviation on one test, combined with failing at -.1 standard deviation on the other two may be more efficient.

The two tests on each of the previous tables having the failure level at -.1 need not be considered individually or when combined. Such consideration has already been made when all three tests had the -.1 failure level. Many of the remaining possible combinations need no investigation because the partial chis involved are very small or are negative.

The most promising combinations are listed in Table 12 with the measures of their ability to fail washouts and non-washouts in different proportion and the measures of their economy in failure.

TABLE 12

COMPARISON OF THE EFFICIENCY OF VARIOUS FAILURE LEVELS

Failure Levels	Chi	% of Washouts Failed	% of Remaining Cadets Failed
Below -.1 on 2 and below -.4 on 6 (without consideration of 4)	4.43	44	14
Below -.1 on 6 or below -.4 on 4 (without consideration of 2)	4.24	91	53
Below -.4 on 2 and below -.1 on 6 (without consideration of 4)	3.97	47	18
Below -.4 on 4 and below -.1 on both 2 and 6	3.83	32	9
Below -.4 on 6 and below -.1 on both 2 and 4	3.73	32	10
Below -1.0 on 2 or below -.1 on 4 (without consideration of 6)	3.53	82	51
Below -.7 on 2 or below -.1 on 6 (without consideration of 4)	3.49	82	51
Below -.7 on 2 and below -.1 on 6 (without consideration of 4)	3.46	32	11
Below -.4 on 2 and below -.1 on both 4 and 6	3.37	32	11
Below -.4 on 6 and below -.1 on 4 (without consideration of 2)	3.34	41	17
Below -1.0 on 4 and below -.1 on 2 (without consideration of 6)	3.33	24	7
Below -.1 on 6 and below -.4 on 4 (without consideration of 2)	3.21	41	18
Below -.4 on 4 and below -.1 on 2 (without consideration of 6)	3.14	41	18

Below -.4σ on Test 2 and below -.1σ on Test 6 (without consideration of Test 4) comes the nearest of any of these combinations to meeting the 50-20 per cent failure requirement. It is close to the failure proportions yielded by failure level at the mean for the same tests. This latter arrangement does meet the criterion.

No other shis and failure percentages show increase in selective efficiency by combining different, instead of the same, cut-off levels on the three tests.

Although these combinations suggest the elimination of the Two-Hand Coordination Test from the battery, it may well be that the present results of this test do not work well in combinations of tests because of the known unreliability of the sampling. All decisions about this test must wait for the evaluation of results obtained from uniform conditions of application.[19]

SECTION II

STATISTICAL TREATMENT OF PART II DATA WITH A COMPARISON OF THE RESULTS FROM PARTS I AND II

This section of the report presents a treatment of Part II data on the most promising tests in the Part I analysis. These tests were selected for further study because they had demonstrated sufficient association with the criterion of pilot success in the first part of the study to justify further investigation.

The same types of analyses are made with the results of the second part of the study as were conducted in the first part. Because the samples used in the two parts differed somewhat,[20] a comparison of the results of both analyses is carried out in all cases.

The following types of statistical treatments of the data are presented in the following pages.[21]

---

[19] Subsequent experimentation conducted by the Committee on Selection and Training of Aircraft Pilots, and more particularly by the Army Air Forces, has borne out the validity of the Two-Hand Coordination Test as a predictor of success in flight training. See McFarland, Ross A, and Chamnell, Ralph C. On. cit. Also Melton, A. W. The selection of pilots by means of psychomotor tests. J. Aviation Med., 1944, 15, 116-123.

[20] McFarland, Ross A. and Fransen, Raymond. On. cit.

[21] Standard scores for all Part II cadets on each test are on file with the National Research Council Committee on Selection and Training of Aircraft Pilots. Also on file are the distributions for each selected test in sigma intervals; distributions of halves of each test in sigma units; distributions of washouts and their theoretical (expected) distribution used in computing the X^2; and distributions of washouts in halves of each selected test.

1. The significance of the difference between the distribution of the scores of washouts for inaptitude and the distribution of scores of all cadets (using chi-squared), and the selection of those tests having a difference that is statistically significant.

2. An evaluation of the reliability (using chi-squared techniques) of the sample by comparison of halves of the population on each of 5 selected tests.

3. Consideration of the efficiency and reliability of various failure levels on these selected tests, i.e., consideration of selected tests in terms to determine the most efficient cut-off points (using chi).

4. Investigation of the various combinations of the selected tests to determine if possible a failure pattern with greater efficiency than that provided by any one test alone (using multiple chi).

COMPARISON OF WASHOUTS FOR INAPTITUDE AND ALL CADETS ON FOURTEEN TESTS OF PART II

Table 13 presents the significance of the difference in the distribution of scores for all cadets and for the washout group (Part II cases). The P-values for this same analysis conducted with the Part I cases are also presented in order to facilitate a comparison of the two groups (Parts I and II) with respect to each of the following tests.

Test Number	Name of Test
1	Eye-Hand Coordination Test
2	Otis Test of Mental Ability
3	Ataxiameter
4	Two-Hand Coordination Test
5	Perception of Change in Position
6	Mashburn Serial Action Test
7	Dotting Test
8	Minnesota Paper Form Board
10	Diastolic Blood Pressure -- reclining
11	Systolic Blood Pressure -- reclining
14	Tidal Air/Body Surface
16	Basal Metabolism Rate
19d	Time Interval to Smallest Pulse Pressure -- Tilt Table
31	Vital Capacity

Comparison with the results of Parts I and II yields first the reassuring fact that Tests 2, 4, and 6, (Otis, Two-Hand, and Mashburn) have the lowest P-values in both samples. The only shifts from reliable to unreliable differentiation or vice versa are in the case of Test 19d (Tilt Table -- time interval

TABLE 13

COMPARISON OF WASHOUTS AND ALL CADETS ON EACH TEST
(N of washouts = 47)

Test Number	Part I P	Part II P	Part II X^2	Part II Degrees of Freedom	Part II N of Parent Distribution
1	.17	.04	13.53	6	146
2	.03	.001	19.98	5	320
3	.80	.68	3.90	6	414
4	.04	.000	34.34	5	362
5	.81	.72	2.85	5	415
6	.0003	.03	10.51	4	395
7	.59	.22	8.22	6	414
8	.44	.05	10.97	5	299
10	.97	.38	4.20	4	180
11	.41	.43	3.85	4	178
14	.91	.80	2.35	5	369
16	.83	.26	6.57	5	362
19d	.06	.48	2.62	3	194
31	.39	.28	6.32	5	342

to smallest pulse pressure) which had borderline validity in Part I and Tests 1 and 8 (Eye-Hand Coordination Test and the Paper Form Board Test) which have borderline validity in Part II.

Three tests, Otis (Test 2), Two-Hand Coordination (Test 4), and the Mashburn (Test 6) are unquestionably able to distinguish between those who were passed and those who were failed in their flight training. It was decided to include Tests 1 and 8 also in the rest of the Part II analysis, even though they evidenced no value in the Part I materials. Assurance of value is confined to Tests 2, 4, and 6. It will be shown that for these tests, statistics obtained from the first group could have been applied to the second with satisfactory predictive efficiency.

It should be noted that in most of the tests the probabilities in the Part II application are appreciably lower than those in Part I. Several explanations are possible. Greater skill may have been acquired in the administration of the tests (i.e., less chance discrepancies in measurements), or a raising of the washout criteria (i.e., failing more strictly on the basis of inaptitude) between the earlier and later pilot classes which might account for the over-all difference. The present interest, however, is in the relative discriminative efficiency of the tests and that remains substantially the same in both samples tested.

COMPARISON OF PARTS I AND II WITH RESPECT TO TESTS 1, 2, 4, 6, AND 8

It would, of course, have been desirable to combine the data from the Part I and Part II samples for the remainder of the analysis since the larger numbers would have increased the reliability of the findings. In order to make this combination, however, it must be demonstrated that the data from the first and second cadet groups are truly representative samples from a homogeneous universe. If such homogeneity is present then the means and the standard deviations of Parts I and II must be alike and the chi-squares[22] of the frequencies throughout the range must be low. These values are presented in Table 14.

None of the tests, with the possible exception of Test 8, shows sufficient homogeneity between the Part I and Part II data. Some of the tests show similar means or similar sigmas or both, and some show similar distributions of their standard scores (i.e., when scores are equated for mean and sigma) but none of them evidence homogeneity throughout. Since each set of standard scores is computed from its own mean and sigma an arbitrary similarity has been introduced. Nevertheless, the P-values are very low.[23] If the standard score distributions of Parts I and II were alike on all tests (high P-values) they might have been combined regardless of differences between means and sigmas. Tests 4 and 8 are the only ones where the differences between the standard score distributions of Parts I and II are small enough to be considered as being due to chance.

Since the two samples cannot be justifiably combined for all five of the tests it seemed advisable to carry through the analyses on Part II separately and compare the results with those obtained from the Part I data.

The question still remains as to why there should be so little similarity between the results of a test when applied to two different groups of pilots who [illegible] differed only by [illegible] time of their entrance into the same training course. This can only be answered in terms of the testing, administration [illegible] real difference in the sample tested or both. [illegible] of selection are a very likely reason [illegible] the difference between groups.

The [illegible] that [illegible] explanation, the samples of [illegible] be considered [illegible] different universes [illegible] they must, therefore, be [illegible] separately [illegible] in [illegible] of the low P-values [illegible] tests (2, [illegible]

---

[22] [illegible] two independent [illegible] chi-square formula [illegible] $\frac{1}{N_1 N_2}$ [illegible] $\frac{(f_1 N_2 - f_2 N_1)^2}{f_1 + f_2}$ [illegible]

[23] [illegible]

TABLE 14

COMPARISON OF PARTS I AND II WITH RESPECT TO THE FIVE SELECTED TESTS

Test Number	P	χ^2	Degrees of Freedom	N of Part II	N of Part I
1	.02	18.18	8	146	373
2	.001	28.29	10	320	336
4	.45	10.89	11	361	371
6	.00000	59.36	7	374	395
8	.28	13.25	11	299	339

	MEAN		SIGMA	
	Part I	Part II	Part I	Part II
1	70.0	71.0	8.7	9.3
2	52.2	45.5	8.7	9.5
4	59.3	67.9	11.9	11.6
6	6.0	5.6	1.0	1.7
8	43.1	43.2	8.4	8.0

experiments. At the same time it aggravates the problem of finding standards or points of cut-off which will be efficient as criteria of selection. Part II results validate the selection value of the three tests. They do not assure any precise danger points.

THE RELIABILITY OF THE SAMPLE

Comparison of the distributions of two halves of each test will indicate the homogeneity within the Part II data. For this purpose the same chi-squared technique is employed as was used in the comparisons of Parts I and II. It will be recalled that the purpose here is not reliability of measurement but the homogeneity of the sample with respect to the test operation.

The Part II measures appear in Table 15 along with comparable measures selected from the analysis of the Part I data. The disparity between the Part I halves of Test 4 was expected because of known lack of calibration in the testing instrument during the testing of the first group. It will be noted that this apparent unreliability in the data of Test 4 disappears in Part II and that all tests show P-values sufficiently large to allow the acceptance of their data as homogeneous and, therefore, susceptible of analysis as single universes of data.

TABLE 15

DISTRIBUTION CHARACTERISTICS OF THE FIVE SELECTED TESTS

Test Number	PART I Comparison of Halves P	PART II Comparison of Halves P	X^2	Degrees of Freedom	PART II Total Distribution Parameters Mean	Sigma	N
1	.22	.12	5.82	3	71.0	9.3	146
2	.88	.59	5.60	7	45.5	9.5	320
4	.06	.55	7.89	9	67.9	11.6	362
6	.73	.29	7.33	6	5.6	1.7	395
8	.79	.78	4.73	8	43.2	8.0	299

RELIABILITY AND EFFICIENCY OF VARIOUS FAILURE LEVELS

Tests 2, 4, and 6 have now been shown to distinguish significantly between successful and unsuccessful cadets on two separate sets of data and it has been found that though the two samples of the experiment cannot be combined, their separate data are homogeneous within themselves on all of these selected tests. In addition, Tests 1 and 8 distinguished between successful and unsuccessful cadets in the Part II analysis sufficiently well to warrant further treatment.

There now arises the problem, specific to pilot selection, of how these tests may be employed to eliminate those cadets lacking flying aptitude from those who were able to attain pilot rank. It will be recalled that this problem is not concerned with the identification of degrees of skill throughout the whole range of flying proficiency but of locating a point in aptitude (as evidenced by performance on certain tests) below which cadets would probably be unable to complete the training program and above which they probably will give at least adequate performance. It is, therefore, a problem of locating a performance level on one or more tests which will predict probable rejection from, or retention in, the pilot training program.

The previous analyses in this section have dealt with differences between the total distributions of test scores for washouts and all cadets. The analysis at this point provides an estimate of the selective efficiency of particular categorical divisions. The same statistical technique that has been used in testing differences between distributions provides an abstract measure of this difference between categorical breaks. Because there are only two classes of each variable (one degree of freedom) chi, not chi-squared, is used but the results have the same interpretation as those previously considered.[14] The various failure levels to be tested by chi must be known to be reliable in the differences they demonstrate. The test for each tested failure level has, therefore, been computed separately for halves of each test. Unless a point of failure reveals a significant difference in high chi on both of the halves, it may not be employed.

CHIS* FOR WASHOUTS AGAINST ALL CADETS AT A SERIES OF FAILURE LEVELS COMPUTED FOR TWO HALVES (A & B) OF EACH SELECTED TEST
(PART II)

Test Number	Halves	Chis for two-place comparison when failure level is at & including: -.3σ	-.1σ	+.1σ	+.4σ	+.1σ
1	A	---	.1[illegible]	1.42	1.29	2.20
1	B	.29	.29	1.40	.52	1.34
2	A	2.21	2.47	2.12	1.32	1.14
2	B	---	3.71	2.09	1.51	1.37
4	A	---	3.25	3.25	3.58	2.66
4	B	---	2.25	2.32	1.97	2.55
6	A	---	---	3.01	3.30	2.14
6	B	---	---	1.37	.38	1.18
8	A	2.40	1.6[illegible]	2.49	1.75	2.32
8	B	1.85	2.5[illegible]	2.16	1.70	1.70

* Chi, not chi-squared, because there is only one degree of freedom in the two-place comparisons.

Various cut-off or failure points on each of the selected tests may be evaluated from the data in Table 16. The chis measure the amount of difference between the proportion of washouts and of all pilots above and below cut-offs placed at every three-tenths of a standard deviation. The chis are directly comparable since there is only one degree of freedom in every comparison. Chis have not been computed for any failure level which would put less than three cases in any category of the comparison.

Tests 4 (Two-Hand Coordination Test) and 8 (Paper Form Board) show very significant and reliable distinctions between successful and unsuccessful pilots at any point below their means. The same is true of Test 6 (Washburn), but less reliably. Test 2 (Otis) makes less and less distinction as it nears the mean and Test 1 (Eye-Hand Coordination Test) might be considered as making adequate differentiation only at -.1 sigma.

The same analysis of the Part I data located the possible cut-offs as presented in Table 17 (insignificant or unreliable chis are not given).

On Tests 2, 4, and 6 several points can be reliably located below which it is probable that there will fall a significantly larger proportion of potential washouts than of potentially successful pilots.

It is necessary next to consider which of the failure points on any test will demonstrate maximum efficiency of elimination. There will be a high chi, for example, if the one-third of the parent population that is separated

TABLE 17

CHIS* FOR WASHOUTS AGAINST ALL CADETS AT A SERIES OF FAILURE LEVELS COMPUTED FOR TWO HALVES (A & B) OF EACH SELECTED TEST (PART 1)

Test Number	Halves	Chis for two-place comparison when failure level is at & including: -1.3σ	-1.0σ	-.7σ	-.4σ	-.1σ
1	A	---	---	---	---	1.07
1	B	---	---	---	---	.93
2	A	1.81	1.11	2.33	2.74	2.33
2	B	1.40	.78	1.14	.96	1.42
4	A	---	1.50	.42	1.66	2.27
4	B	---	1.82	2.40	2.66	2.33
6	A	---	---	1.14	1.22	2.10
6	B	---	---	2.29	3.96	3.01
8	A	---	---	3.02	1.80	1.75
8	B	---	---	-.90	.41	.49

* Chi, not chi-squared, because there is only one degree of freedom in the two-place comparisons.

by a given cut-off includes two-thirds of the criterion group. But in the problem of pilot selection sacrificing one-third of potential pilot material in order to eliminate two-thirds of the potential washouts is very costly. In the present situation, with the need for pilots great and training facilities limited in terms of that need, it is essential to find a cut-off that will eliminate a maximum of washouts and a minimum of those who in all probability will attain pilot rank.

Table 18 shows the per cent of washouts (W.O.) and of remaining cadets (R.C.) who would be failed by each of the tests at each of the failure levels tested for significance by the chis of the preceding table.[25]

Test 2 (Otis) at one sigma below the mean fails almost half of the washouts and only one-seventh of the remaining cadets in this sample. If these figures could be relied upon to hold throughout the total universe

---

[25]Up to this point the washouts have been compared with the total pilot sample. That is, they have been differentiated from a "normal" pilot population that included the unsuccessful as well as the successful. From this point on all comparisons are between the washouts and those who presumably are still accepted pilots.

TABLE 18

PER CENT OF WASHOUTS AND OF REMAINING CADETS WHO WOULD BE ELIMINATED BY EACH FAILURE LEVEL
(PART II)

Test Number	Failure Level at and Including: -1.3 sigma W.O. %	-1.3 sigma R.C. %	-1.0 sigma W.O. %	-1.0 sigma R.C. %	-.7 sigma W.O. %	-.7 sigma R.C. %	-.4 sigma W.O. %	-.4 sigma R.C. %	-.1 sigma W.O. %	-.1 sigma R.C. %
1			15	14	36	18	47	33	66	38
2	34	8	43	14	47	25	53	37	66	51
4			38	14	53	24	68	36	79	48
6					34	15	45	26	64	45
8	30	12	38	17	47	22	55	35	66	42

of prospective pilots selection might be based on the Otis Test alone. In the Part I data, however, the Otis did not show such a high degree of selective efficiency. In Table 19 the percentages of elimination in both parts are presented in terms of the most efficient failure levels found in the Part II sample.

Although judging from percentages of the second set of data it would seem satisfactory to use Test 2 or Test 4 at -1.0 Sigma, it is apparent, when viewing both sets of data, that these percentages are not dependable. It is evident from the recurrence of high chi values at the same failure levels for the same tests in both sets of data, that any one of the tests can be relied upon to make the desired differentiation. However, the efficiency of this selection cannot be trusted. Without further evidence, no one of the tests alone has a dependable economy in this differentiation. This is due partly to the size of the washouts sample and partly to the demonstrated difference in the character of the Part I and Part II samples.

TABLE 19

THE MOST EFFICIENT SELECTIONS BASED ON PART II

Test Number	Failure Level at and Including: -1.0 sigma Part II W.O. %	-1.0 sigma Part II R.C. %	-1.0 sigma Part I W.O. %	-1.0 sigma Part I R.C. %	-.7 sigma Part II W.O. %	-.7 sigma Part II R.C. %	-.7 sigma Part I W.O. %	-.7 sigma Part I R.C. %
2	43	14	27	17	---	---	---	---
4	38	14	29	13	---	---	---	---
6	---	---	---	---	34	15	40	21

There can be no doubt that extremely low intelligence and very poor psychomotor coordination (as measured by these tests) are closely associated with inability to learn to fly. The presence of either one alone, however, is not a sufficient guarantee of pilot failure to be of practical use. The effect of these handicaps in combination must still be investigated in combination.

FAILURE PATTERNS FROM TESTS 2, 4, 6, AND 8 IN COMBINATION

If high intelligence (high relative to the group under consideration) can compensate for poor psychomotor ability (also relative to the group) then failure on either measure alone will not be as closely associated with inability to fly as failure on both.

In order to investigate how the tests in combination operate to distinguish washouts from successful pilots, the same chi technique employed throughout is applied in the form of partial and multiple chis.[26] A partial chi is a measure of how "fail" (at any given definition of failure) on any one test with "pass" on the other tests is associated with the criterion (being a washout). It may also be obtained for the cases who "fail" on two or three tests and "pass" the other two or one tests.

In the process of computing it is unnecessary to carry these partial measures through to the final chi, for the delta used in the formula for chi,

$$\left(\frac{\Delta^2 N}{(a+b)\ (c+d)\ (a+c)\ (b+d)}\right)^{1/2}$$

is the determinant of the chi's value and may be used directly in estimating the worth of any "pass-fail" combination.[27] These values (deltas) appear in Table 21.

From the deltas in Table 20 the contribution of any test may be analyzed independent of the other tests toward predicting the criterion of flying failure.[28] A negative delta means that the differentiation is in the unexpected direction, i.e., there are fewer washouts and more remaining cadets selected by the given "pass-fail" dichotomy than would be expected from the total distribution of pilots.

---

[26] Franzen, Raymond. On. cit.

[27] $\Delta = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$ A discussion of this short-cut procedure is to be found in: Franzen, Raymond. On. cit. Appendix A.

[28] Test 1 (Eye-Hand Coordination Test) has been omitted from the analysis of the tests in combination. This analysis requires, of course, that all cases have scores on all tests. Very few of the cadets who had scores on the four other tests also had scores on Test 1. Inclusion of the test would have seriously reduced the size of the sample.

TABLE 20

DETERMINANTS USED IN SELECTING HIGHEST CHIS FOR FAILURE ON COMBINATIONS OF TESTS AT VARIOUS FAILURE LEVELS

Pass & fail permutations* Tests 2 4 6 8	Failure level = -.1 sigma on all four tests			Failure level = -.4 sigma on all four tests			Failure level = -.1 on Tests 4 & 8 and -.4 on Tests 2 & 6		
	No. failed			No. failed			No. failed		
	W.O.	R.C.	Delta	W.O.	R.C.	Delta	W.O.	R.C.	Delta
- + + +	2	15	-351	2	13	-257	1	11	-340
+ - + +	2	11	-163	2	15	-351	3	21	-456
+ + - +	-	12	-564	1	7	-152	1	6	-105
+ + + -	2	12	-210	4	14	50	3	19	-362
- - + +	1	9	-246	2	10	-116	2	9	-69
- + - +	2	8	-22	-	4	-188	-	3	-141
- + + -	2	8	-22	3	12	-33	2	9	-69
+ - - +	6	11	545	5	6	603	5	6	603
+ - + -	1	5	-58	3	8	155	5	11	368
+ + - -	1	4	-11	1	5	-58	1	1	130
- - - +	3	7	202	4	1	661	2	2	260
- - + -	7	9	816	6	9	639	8	15	711
+ - - -	4	10	238	2	5	119	2	10	-116
- + - -	1	9	-246	-	6	-282	-	4	-188
- - - -	13	23	1220	8	9	993	10	11	1253
Total N	47	177		47	177		47	177	

* + = scores above failure level on that test
\- = scores at and below failure level on that test

The hypothesis of compensation is clearly demonstrated by the first four large negative deltas which measure the predictive efficiency of failure on each of the four tests plus passing the other three tests. In other words, if a candidate is low in the ability measured by any one of the tests, but high in the abilities measured by the other three, he has a better than average chance of becoming a successful pilot. It is not demonstrated in the Part II materials that intelligence and psychomotor coordination are completely compensatory, though this seemed to be the case in Part I. Test 8, the Minnesota Paper Form Board, complicates the interpretation of the deltas since it is a "special aptitude" test and cannot be classified with either the Otis or the other two that are psychomotor tests. Nevertheless the combination (pass on Test 2 with fail on 4, 6, and 8) and the combination (pass on Test 2 and 8 with fail on 4 and 6) both provide efficient differentiation of washouts. If intelligence or intelligence plus the special aptitude definitely compensated for poor psychomotor functions, the deltas of these combinations would be very low or negative.

That some compensation exists among all four tests is unquestionable, not only because of the negative deltas for failure on only one of the tests, but because of the very high positive delta when the standard is failure on all four.

The highest deltas are obtained from the failure patterns that include Tests 4 and 6 or Tests 4 and 8. The significance of Test 4 in Part II materials is further evident in the sizable negative delta for the combination. However, there are only ten cases in this pattern.

The Part I data indicated compensation between Tests 2 and 6. Test 4 in those materials was known to be unreliable which may explain the absence of the Tests 4 and 6 compensation found here. If Tests 2 and 4 were correlated the different types of compensation found in the two sets of materials would be readily understandable.

Intercorrelations of all of Tests 1 to 9 (Table 2) were obtained for the Part I materials and were found to be very low throughout. Using a contingency technique the relation between Tests 2 and 4 is found to be very low in Part II even though Test 4 materials have been purified and found reliable:

	Test 2	
Test 4	Pass	Fail
Pass	55	47
Fail	50	49
	Chi = .51	

The chis and efficiency percentages of the various combinations of any two, three, or four of these tests may now be examined. The partial deltas show the predictive efficiency of failure on any test or tests even though the other tests are passed. Failure on a particular test or tests is thereby isolated from failure on the other tests. In contrast, the chis of Table 21 measure the value of combinations of tests irrespective of the other tests (i.e., as if we had no data on the tests not in the combination).

All of these chis are large enough to indicate that the tests in combination have a very significant function in distinguishing washouts.

Comparison of the chis obtained in the two samples indicates changes in chi value but not in the validity of [illegible] tested combinations in both parts. Table 21 presents comparable statistics on Parts I & II and Table 22 presents all [illegible] in Part [illegible].

[illegible] of these tests when they [illegible] of successful and [illegible] unsatisfactory. Comparison [illegible] in chis and cost [illegible] of elimination.

CHI SQUARE TEST OF ASSOCIATION BETWEEN PAIRS OF TESTS

Failure Criteria*	Part I: Per cent failed			Part II: Per cent failed		
	Chi	W.O.	R.C.	Chi	W.O.	R.C.
Failure level = -.4 sigma on all tests						
4 & 6	3.40	55	29	3.46	50	23
2 & 4	3.12	51	27	2.75	47	25
2, 4, & 6	2.57	34	17	3.87	38	13
2 & 6	1.85	40	27	4.33	53	20
Failure level = -.4 sigma on all tests						
4 & 6 -	4.54	40	12	3.39	35	13
2, 4, & 6	4.07	26	6	3.33	24	7
2 & 4	3.86	43	16	1.99	29	16
2 & 6	2.48	26	11	4.39	41	12
Failure level = -.1 sigma on Test 4 and -.4 sigma on Tests 2 & 6						
2 & 4	3.50	47	21	3.14	41	18
4 & 6	3.57	40	26	3.34	41	17
2, 4, & 6	3.52	26	7	3.37	32	11
2 & 6	**					

* Taken from Table 11, page [illegible].

** This combination is -.4 sigma on Tests 2 and 6 and appears in the section above.

Parts I and II have been treated separately first because of the differences in their distributions on the selected tests and second because it seemed advisable to test the reliability of the Part I findings. The final failure patterns of the Tests 2, 4, and 6 in combination show surprising consistency between the two samples studied. The sample may therefore be justifiably combined in these dichotomous arrangements determined by their respective standard deviations. The combined results are given in Table 23.

It will be noted that the chis are all increased in Table 23. This increase is due to the larger number of cases involved in the comparisons. In the third failure level, using different cut-offs on the different tests, the Test 6 level has been placed at -.1 sigma instead of -.4 as in Part II because Part I data showed this to be a more effective break.

In all but one of the pairs of percentages the washout proportion is between two and three times as large as the proportion of non-washouts. The test combinations of Tests 4 and 6, and 2 and 4 with failure set approxi-

TABLE 22

CHIS FOR FAILURE ON COMBINATIONS OF TESTS
(PART II)

Failure on tests:	Chi	Per cent failed W.O.	Per cent failed R.C.
Failure level = -.1 sigma on all 4 tests			
2, 4, & 8	3.53	43	18
4 & 8	3.48	53	27
4 & 6	3.40	55	29
2 & 4	3.12	51	27
2, 4, & 6	2.57	34	17
4, 6, & 8	2.41	36	19
2, 4, 6, & 8	2.15	27	13
2 & 6	1.85	40	27
Failure level = -.4 sigma on all 4 tests			
4 & 6	4.54	40	12
2, 4, & 6	4.07	26	6
2 & 4	3.86	43	16
2, 4 & 8	3.42	30	10
4 & 8	3.35	40	18
2, 4, 6 & 8	2.75	17	5
4, 6, & 8	2.63	21	8
2 & 6	2.48	26	11
Failure level = -.1 on Tests 4 & 8 and -.4 on Tests 2 & 6			
2, 4, 6 & 8	4.42	21	6
2, 4, & 8	3.61	38	15
2 & 4	3.58	47	21
4 & 6	3.57	40	16
2, 4, & 6	3.52	26	7
4, 6, & 8	2.35	26	12

mately at the mean, eliminates half the washouts and one-fourth of the remaining cadets. If such a proportion of unnecessary rejections is too extravagant, Tests 2, 4, and 6, with failure at -.4 sigma, will eliminate one-fourth of the potential washouts and only six per cent of the potential pilots. Tests 2 and 6 in combination operate best when -.1 sigma is the cut-off for 6 and -.4 sigma for 2. Then the rejected group includes two-fifths of the washouts and only one-fifth of the desirable candidates.

Two combinations may be used as alternatives. If failure on Tests 4 and 6 or failure on Tests 2 and 4 is used as a standard, 67 per cent of the washouts and 36 per cent of the remaining cadets are eliminated.

TABLE 23

CHIS FOR FAILURE ON COMBINATIONS OF TESTS
(PART I and PART II combined)

Failure on tests:	Chi	Per cent failed W.O.	Per cent failed R.G.
Failure level = -.1 sigma on all 3 tests			
4 & 6	5.17	53	25
2 & 4	5.05	49	26
2, 4, & 6	4.70	36	14
2 & 6	4.43	46	22
Failure level = -.4 sigma on all 3 tests			
4 & 6	5.75	38	13
2, 4, & 6	5.33	25	6
2 & 6	4.68	32	12
2 & 4	4.44	37	16
Failure level = -.1 sigma on Tests 4 & 6 and -.4 on Test 2			
2, 4, & 6	4.61	32	12
2 & 4	4.23	43	21
2 & 6	4.12	40	19

That combinations of tests can operate reliably and significantly to differentiate poor pilot material and can operate better than they do separately seems incontrovertible. Further experiments with other tests such as Tests 1 and 8 and the inclusion of reliable cardio-vascular and respiratory measures should reveal a combination which would reject with greater precision so that there need not be the unwanted loss of good pilot material.

TENTATIVE CONCLUSIONS FROM THE ANALYSIS OF THE DATA IN SECTIONS I AND II

The data in Sections I and II of this study were subjected to the same types of statistical treatment, with an additional discussion in Section II of a comparison of the results of the two independent analyses. Findings have suggested the following:

The Otis and Washburn Tests, and, provisionally, the Two-Hand Coordination Tests, have been selected as the predictive battery because of their ability to measure characteristics that occur differently among washouts than among all cadets.

A level of selective efficiency for the tests set at below average on both the Otis and Washburn, or on all three tests, will yield a large difference between proportions of washouts and proportion of non-washouts failed by the tests.

The proportions so failed are efficient and economical, being about half of the washouts and not more than one-fifth of the non-washouts.

One important advantage gained from the application of the multiple cut technique is the possibility of conclusions with reference to the compensation relation among these tests. Knowing that these tests are all related to the probability of being washed out, does not tell whether failure in general intelligence is compensated by success in the psychomotor examinations, or vice versa. When the analysis described in the previous pages was applied it was found that such compensation did exist, i.e., being low in psychomotor ability, but above average intelligence does not predict rejections nearly as well as being low in both traits. Likewise it can be seen 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.