ON THE ACTUAL AND POTENTIAL VALUE OF BIOGRAPHICAL INFORMATION AS A MEANS OF PREDICTING SUCCESS IN AERONAUTICAL TRAINING

by

H. M. Johnson

In cooperation with

Mary L. Boots and Robert J. Wherry

with the assistance of

Olive C. Hotaling
Lillian Holt Martin
Frank P. Cassens, Jr.

A report on research conducted at Tulane University and the University of North Carolina with data gathered at the Naval Air Station, Pensacola, Florida, in cooperation with the Bureau of Aeronautics of the U. S. Navy, 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.

[illegible] ADMINISTRATION
[illegible]

9-1-44

National Research Council

Committee on Selection and Training of Aircraft Pilots

Executive Subcommittee

M. S. Viteles, Chairman

C. W. Bray

D. R. Brimhall

L. A. Carmichael

J. W. Dunlap

J. C. Flanagan

H. M. Johnson

W. R. Miles

G. R. Wendt

National Research Council

1944

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

August 21, 1944

Dr. Dean R. Brimhall
Director, Airman Development Division
Civil Aeronautics Administration
Washington 25, D. C.

Dear Dr. Brimhall:

Attached is a report entitled On the Actual and Potential Value of Biographical Information as a Means of Predicting Success in Aeronautical Training and Rate of Learning prepared by H. M. Johnson, in cooperation with Mary L. Boots and Robert J. Wherry and with the assistance of Olive C. Hotaling, Lillian Galt Martin, and Frank P. Cassens, Jr. This report is submitted by the Committee on Selection and Training of Aircraft Pilots with tho recommendation that it be included in the series of technical reports issued by the Airman Development Division, Civil Aeronautics Administration.

The report describes a study of biographical items in predicting performance during flight training conducted at the U. S. Naval Air Station, Pensacola, Florida. The investigation was made possible through the cooperation of the Bureau of Medicine and Surgery, U. S. Navy, and of officers attached to the U. S. Naval Air Station at Pensacola, Florida.

The report is one of a series growing out of the Pensacola Project. It is of particular significance in revealing the effectiveness of biographical items, as contrasted with the more commonly used psychological tests, in the prediction of flight proficiency. It is of interest to note that this investigation, and other studies of biographical data conducted under the auspices of the Committee on Selection and Training of Aircraft Pilots have had very practical outcomes in the development of materials actually used by the U. S. Navy and in other programs for the selection of pilots.

Cordially yours,

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

MSV:ts

ACKNOWLEDGMENTS

The undersigned assumes responsibility for the accuracy of description of this study and of its results. But for obtaining various sections of them responsibility belongs to the various people mentioned throughout the report. If time had permitted they would have been asked to edit and co-sign the report, but time presses upon all and some of the naval officers are on sea duty, and other civilian assistants are engaged in other work. Special acknowledgments belong to Captain Frederick Ceres (MC) USN, Medical Officer, Captain Albert C. Read, USN, Commandant, and Commander (now Captain) W. D. Sample, USN, Executive Officer, all at Pensacola, for granting special courtesies and facilities that made the work possible. It is a pleasure to acknowledge also the valuable counsel of Lieut. R. H. Peckham, H(V)-S, USNR, and of the extraordinarily faithful, diligent, and accurate work of assistants: namely, Olive C. Hotaling, Lillian Galt Martin, and Frank Paul Cassens, Jr.

H. M. Johnson
Professor of Psychology
Tulane University

EDITORIAL FOREWORD

One major outcome of research on selection techniques conducted by the Committee on Selection and Training of Aircraft Pilots has been the development of a Biographical Inventory which has been used extensively by the U. S. Navy and in the pilot screening program administered by the Committee for the C.A.A. under the title "C.A.A. National Testing Service."

The instrument finally used for this prupose was the immediate outgrowth of work done by E. L. Kelly, of Purdue University. However, this effort was preceded by considerable research involving the analysis of biographical information by H. M. Johnson and his staff at Tulane University with the cooperation of R. J. Wherry and his staff at the University of North Carolina. This study, involving the use of data taken from medical and training records at the U. S. Naval Air Station, Pensacola, Florida, was made possible only through the cooperation of the U. S. Navy.

The report by H. M. Johnson and his associates, dealing with the analysis of biographical information, is one of a series growing out of the Pensacola Project in which it was possible to study not only biographical information but also the relationship between many other psychological and physiological predictors and success in learning to fly.

CONTENTS

[illegible]

1. This is a report of a study of the value of intercorrelated items of biographical information, compiled from medical and training records of 480 students at the U. S. Naval Air Station, Pensacola, belonging to classes 80-83, 86-88, and 90, as predictors of eventual success or failure in the course, and also as predictors of velocity in learning.

2. The study was limited to these classes because only among them was certain important information available. This consisted (1) of 125 items in Bernreuter's personality tests, obtained at the instigation of Lt. Cmdr. Rex H. White (MC) USNR and at his own expense; and (2) of results of systematic personal interviews planned by Cmdr. Victor S. Armstrong (MC) USN and Lt. Cmdr. Wilbur E. Kellum (MC) USN, together with Dr. White.

3. The information was collected while these students were at the station by officers of the medical department and the training department without regard to its usefulness in this connection. In a word, therefore, this is a planned statistical study of records which were not originally intended for use according to this plan. This fact accounts for some items being omitted on some students, so that on only 208 students out of 480 could every item be obtained. The data are therefore exhibited in two sets of tables: namely, (1) Tables 1-A and 1-B which include only completed records; and (2) Tables 2-A and 2-B, in which each student for whom any item of information was lacking was credited with the average score which belonged to that item. The tables and graphs included in this summary show how important it is to get complete records if maximum accuracy of prediction is to be obtained.

4. A description of the individual items and of the mode of combining is included in the main report. This summary includes only the results. Those items which we have considered are all those and only those which could have been collected during the first week of the student's residence at the station. If certain ground school courses, such as practical navigation, had been required before the candidates entered the station, the students' progress in those courses would have been a valuable addition to the items used, and would have increased the predictive value of the whole group of items materially.

5. From Tables 2-A and 2-B, one notes that if the critical score had been set at 0.6, then of the 480 applicants, 294, constituting 61.25 per cent would have been retained, and 186 applicants, constituting 38.75 per cent, would have been rejected. Among the 294 applicants who would have been retained would have been included 240 passers, constituting 81.63 per cent (instead of 64.17 per cent) of those retained, and 54 failers, constituting 18.37 per cent (instead of 35.83 per cent) of those

TABLE 1-A

EFFECT OF APPLYING VARIOUS CRITICAL TEST SCORES
(Students having complete records only)

	RETAINED		REJECTED		PASSERS RETAINED		FAILERS RETAINED		PASSERS REJECTED		FAILERS REJECTED	
Critical test score	Number	Per cent of total	Number	Per cent of total	Number	Per cent of passers	Number	Per cent of failers	Number	Per cent of passers	Number	Per cent of failers
1.2	0	0.00	208	100.00	0	0.00	0	0.00	127	100.00	81	100.00
1.1	2	0.96	206	99.04	2	1.57	0	0.00	125	98.43	81	100.00
1.0	11	5.29	197	94.71	10	7.87	1	0.12	117	92.13	80	98.77
0.9	24	11.54	184	88.46	23	18.11	1	0.12	104	81.89	80	98.77
0.8	56	26.92	152	73.08	47	37.01	9	11.11	80	62.99	72	88.89
0.7	90	43.27	118	56.73	78	61.42	12	14.81	49	38.58	69	85.19
0.6	117	56.25	91	43.75	98	77.17	12	23.46	29	22.83	62	76.54
0.5	142	68.27	66	31.73	110	86.61	32	39.51	17	13.39	49	60.49
0.4	174	83.65	34	16.35	121	95.28	53	65.43	6	4.72	28	34.57
0.3	186	89.42	22	10.58	123	96.85	63	77.78	4	3.15	18	22.22
0.2	197	94.71	11	5.29	125	98.43	72	88.89	2	1.57	8	11.11
0.1	206	99.04	2	0.96	127	100.00	79	97.53	0	0.00	2	2.47
0.0	207	99.52	1	0.48	127	100.00	80	98.77	0	0.00	1	1.23
-0.1	208	100.00	0	0.00	127	100.00	81	100.00	0	0.00	0	0.00

TABLE 1-B

COMPOSITION OF GROUPS RETAINED BY VARIOUS CRITICAL TEST SCORES
(Students having complete records only)

	PASSERS OR FAILERS		PASSERS		FAILERS	
Critical test score	Total number retained	Per cent of total number retained	Number retained	Per cent number retained	Number retained	Per cent number retained
1.2	0	0.00	0	-	0	-
1.1	2	0.96	2	100.00	0	0.0
1.0	11	5.29	10	90.91	1	9.09
0.9	24	11.54	23	95.83	1	4.17
0.8	56	26.92	47	83.93	9	16.07
0.7	90	43.27	78	86.67	12	13.33
0.6	117	56.25	98	83.76	19	16.24
0.5	142	68.27	110	77.46	32	22.54
0.4	174	83.65	121	69.54	53	30.46
0.3	186	89.42	123	66.13	63	33.87
0.2	197	94.72	125	63.45	72	36.55
0.1	206	99.04	127	61.65	79	38.35
0.0	207	99.52	[illegible]	61.35	[illegible]	38.65

TABLE 2-A

EFFECT OF APPLYING VARIOUS CRITICAL TEST SCORES
(All students)

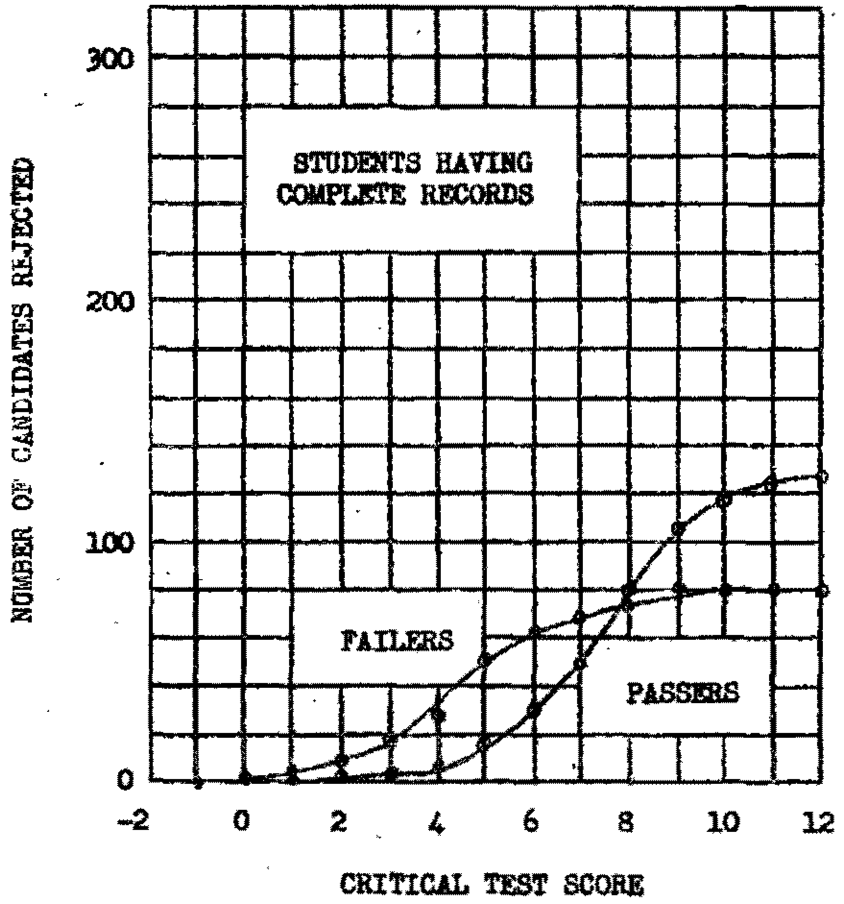
	RETAINED		REJECTED		PASSERS RETAINED		FAILERS RETAINED		PASSERS REJECTED		FAILERS REJECTED	
Critical test score	Number	Per cent of total	Number	Per cent of total	Number	Per cent of passers	Number	Per cent of failers	Number	Per cent of passers	Number	Per cent of failers
1.2	0	0.00	480	100.00	0	0.00	0	0.00	308	100.00	172	100.00
1.1	4	0.83	476	99.17	4	1.30	0	0.00	304	98.70	172	100.00
1.0	13	2.71	467	97.29	12	38.96	1	0.58	296	96.10	171	99.42
0.9	38	7.92	442	92.08	36	11.69	2	1.16	272	88.31	170	98.84
0.8	88	18.33	392	81.67	78	25.32	10	5.81	230	74.68	162	94.19
0.7	219	45.62	261	54.38	188	61.04	31	18.02	120	38.96	141	81.98
0.6	294	61.25	186	38.75	240	77.92	54	31.40	68	22.08	118	68.60
0.5	366	76.25	114	23.75	277	89.94	89	51.74	31	10.06	83	48.26
0.4	421	87.71	59	12.29	296	96.10	125	72.67	12	3.90	47	27.33
0.3	445	92.71	35	7.29	302	98.05	143	83.14	6	1.95	29	16.86
0.2	465	96.88	15	3.13	306	99.35	159	92.44	2	0.65	13	7.56
0.1	477	99.38	3	0.63	308	100.00	169	98.26	0	0.00	1	0.58
0.0	479	99.79	1	0.21	308	100.00	171	99.42	0	0.00	1	0.58
-0.1	480	100.00	0	0.00	308	100.00	172	100.00	0	0.00	0	0.00

TABLE 2-B

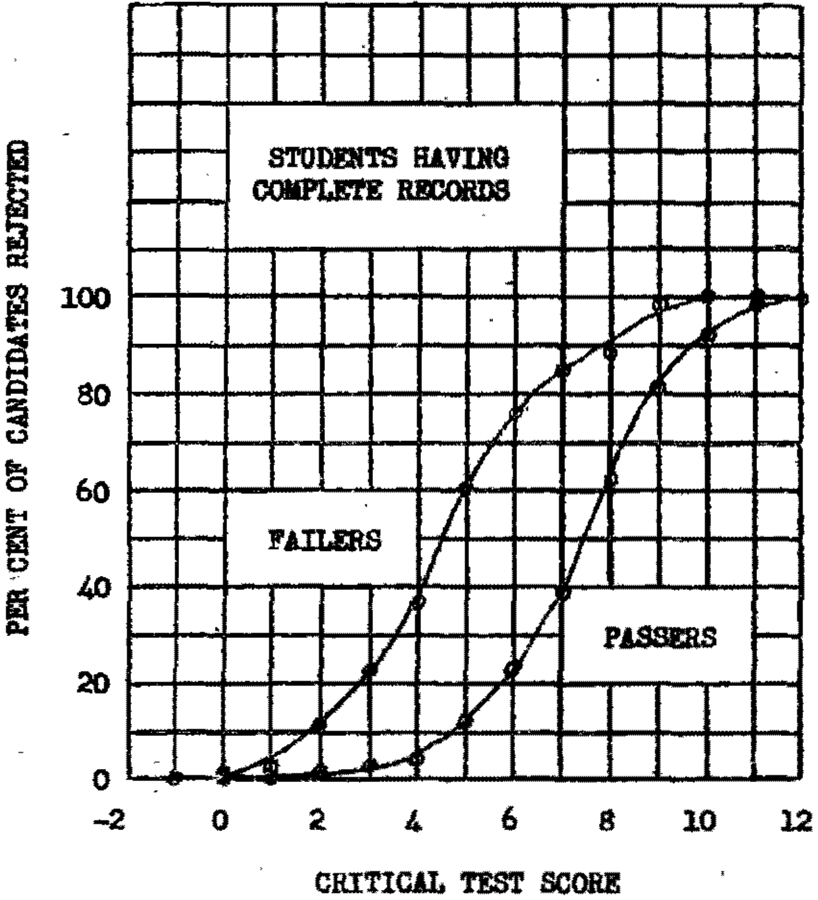
COMPOSITION OF GROUPS RETAINED
BY VARIOUS CRITICAL TEST SCORES
(All students)

	PASSERS OR FAILERS		PASSERS		FAILERS	
Critical test score	Total number retained	Per cent of total number retained	Number retained	Per cent number retained	Number retained	Per cent number retained
1.2	0	0.00	0	-	0	-
1.1	4	0.83	4	100.00	0	-
1.0	13	2.71	12	92.31	1	7.69
0.9	38	7.92	36	94.74	2	5.26
0.8	88	18.33	78	88.64	10	11.36
0.7	219	45.62	188	85.84	31	14.16
0.6	294	61.23	240	81.63	54	18.37
0.5	366	76.25	277	75.68	89	24.32
0.4	421	87.71	296	70.31	125	29.69
0.3	445	92.71	302	67.87	143	32.13
0.2	465	96.88	306	65.81	159	34.19
0.1	477	99.38	308	64.57	169	35.43
0.0	479	99.79	308	64.30	171	35.70
-0.1	480	100.00	308	64.17	172	35.83

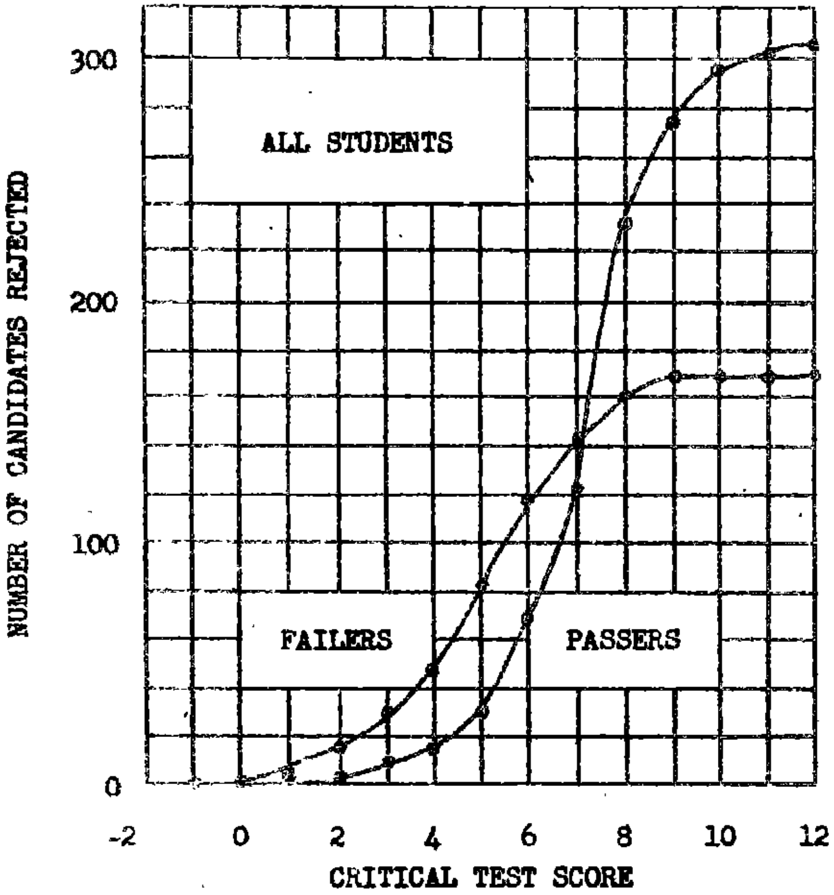
GRAPH 1-A



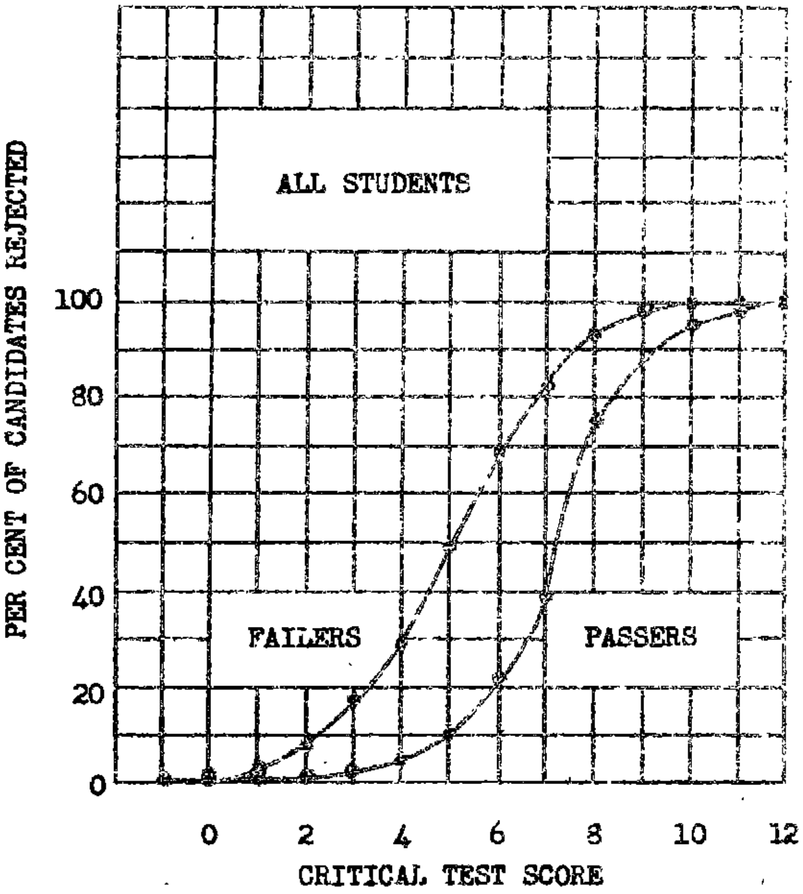
GRAPH 1-B



GRAPH 2-A



GRAPH 2-B



retained. Thus the proportion of failers in the group retained would have been about half the proportion in the original sample. Whether or not it would have been expedient to set the critical score at 0.6, and thus achieve this result, would depend on the number of candidates available. If the 186 candidates rejected by the test could be replaced from the original reservoir, it would unquestionably pay to set the critical score at least as high as 0.6. In fact, if as many as 261 additional candidates were available, it would pay to set the critical score as high as 0.7, for as is shown in Table 2-B, the group of students satisfying this critical score would be made up of 85.84 per cent passers (instead of 64.17 per cent) and 14.16 per cent failers (instead of 35.83 per cent). In general the usefulness of a selective procedure depends on the number of men available to choose from: the larger the reservoir, the more potential passers one can profitably sacrifice in order to reduce the number of failers. In these two tables, as well as in Graphs 2-A and 2-B, the general result of choosing one critical test score rather than another is evident.

6. Tables 1-A and 1-B and the corresponding Graphs 1-A and 1-B deal with only 208 students for whom all the biographical information was available. It is evident from inspection that the predictive value of the complete battery of tests is much greater than if a score on one or more omitted items is replaced by the average score attained by the group. From these tables and graphs it is evident that if the critical score were set at 0.6, 43.75 per cent of the applicants would have been rejected; but of those who were retained, 83.76 (instead of 61.06 per cent) would have been passers and 16.24 per cent (instead of 38.94 per cent) would have been failers. Again, the best score to choose as critical depends on the ratio of the number of candidates who are available to the number of candidates who are needed.

7. In this sample, 308 students, constituting 64.17 per cent of the 480 entrants, completed the course. The selective power of the tests varies according to the proportion of passers. In the main body of the report is presented and proved a generalized formula for predicting the selective power of a test when the population on which it is to be used is not homogeneous with the sample on which it has been standardized.

8. The same items of information, being properly weighted for another purpose, were drawn upon for predicting the time investment required by the various passers for completing the course at Pensacola. The value of the tests in selecting out the slow learners was found to be considerably less than for selecting potential failers.

9. In particular, the analysis disclosed that there is only a very slight tendency for students who require extra time in the earlier stages to require more extra time than the average students in later stages. This finding seems to confirm the wisdom of the present policy of granting extra time to individual students in the stages which they find most difficult.

10. Although very few students were failed because of their records in ground school, nevertheless, the number of ground school courses which the student had to repeat in order to satisfy this requirement, proved to have great predictive value. However, it could not be used in this instance because the ground school course was not given until the students entered the station, and in many instances it was not completed until the student had passed the stage of training before which nearly all the failers in the flying courses are eliminated. However, if these courses are hereafter to be given at the air bases, this relationship can be utilized to advantage. In fact, it seems safe to expect that if the courses in radio and practical navigation are given at the bases, and the hours of flying instruction at the bases are increased to 33, and if such selection devices are employed as have been tested in this study and others of the same kind, then the proportion of washouts at the naval stations need not exceed five per cent.

11. Recommendation: In various informal progress reports the undersigned remarked that the items of biographical information which we studied were not selected according to a rigid plan based on extensive experience. Rather, they were made up of such questions as occurred to some intelligent medical officers who had had some experience in aeronautics. I recommended that an enlarged questionnaire be drawn up which should include those items in this study which proved to have diagnostic significanse along with such other items as the members of the Subcommittee and their colleagues might judge to be suitable for inclusion. It was also pointed out that the significant items taken from the Bernreuter test had been removed from their former matrix and that some of their significance might be lost. It was suggested that at least some of the preceding questions be included with them in order to lessen this effect. These suggestions have already been carried out partly in the Kelly blank and partly in the enlarged biographical inventory that has been used by the U. S. Navy and by the C.A.A.-National Testing Service as administered by the Committee on Selection and Training of Aircraft Pilots.[1]

12. The detailed description of the biographical items, the manner in which they were selected, and the mode of combining them, is set forth along with an historical background in the main body of the report.

[1] Editor's Note Representing modifications of the Biographical Inventory originally prepared in research by E. L. Kelly at Purdue University.

PART II: DETAILED DESCRIPTION

Genesis of the study

Certain medical officers at the U. S. Naval Air Station at Pensacola interviewed 990 students in classes 25-44, and from their personal impressions, somewhat liberally categorized, they predicted whether the student would complete the training course or be dropped. For our present purposes, we shall consider the completers as passers and the non-completers as failers, whether the failure to complete was due to flying ineptitude or not. This mode of classification is at least unambiguous. It has been adversely criticized on the ground that a small proportion of the students drop out for non-flying reasons. Among them are uncontrollable airsickness, effects of acute illness, physical defects undisclosed at the physical examination, accidental injuries, unauthorized marriage, insubordination, indications of moral, mental, or social unfitness for the duties of an officer, etc. None of these non-flying reasons is often mentioned, although one may suppose that if any is present, a check pilot might therefore judge the student's flying work more severely than if it was absent. However, since the purpose of the selective examinations is to admit a maximum number of candidates who will pass and a minimum number who will fail, if a composite selection device fails to select against candidates with any predisposing weakness, it is in so far unsatisfactory. In the list which we have just mentioned, acute illness, especially if infectious and not due to personal carelessness, is the only reason which we should regard as being non-controllable and therefore non-predictable. It has been urged, for example, that if a cadet marries contrary to his agreement with the government, the fact has no relevancy to his qualifications as an aviator. It may in fact not be relevant to his ability to handle a ship. On the other hand he is being trained for a wider range of duties than merely handling the controls of a plane. If he makes an agreement and then intentionally violates it, the fact is certainly relevant to his fitness to exercise command. Hence, if a selection procedure were perfect, it should be sensitive to such personal weaknesses as this, however amiable they may be if manifested in persons who have not obligated themselves.

Without describing the interview in detail, we may mention that it was partly formal and partly informal, that the choice of questions to be asked was left largely to the examiner, and that the over-all rating of the candidate seems not to have been proscribed by a set of unequivocal and rigid rules. Moreover, there is evidence in the records that in at least some of these classes, the subjects were not interviewed until they had reached a stage of training before which most of the complaints and failures occur; and that, therefore, it would be hard to dissociate the impressions of the medical examiner obtained in the interview from his knowledge of those other facts. How important this feature of the situation may have been cannot be determined from the assemblage of data which we shall consider.

Summary of initial results

These 20 classes included 990 students who were rated by the medical examiners and who have now failed or passed the course. The findings can be expressed in a 2 x 2 table, as follows:

TABLE 3

	A = predicted to pass	A' = predicted to fail	Total
B = passed	f = 274	f = 282	556
	f_o = (217)	f_o = (339)	(556)
	Δ = 57	Δ = -57	0
	Δ^2 = 3249	Δ^2 = 3249	-
	x^2_e = 14.97	x^2_e = 9.58	-
B' = failed	f = 113	f = 321	434
	f_o = (170)	f_o = (264)	(434)
	Δ = -57	Δ = 57	0
	Δ^2 = 3249	Δ^2 = 3249	-
	x^2_e = 19.11	x^2_e = 12.31	-
Total	f = 387	f = 603	990
	f_o = (387)	f_o = (603)	(990)
	Δ = 0	Δ = 0	0

x^2 = 55.97, $P = 3(10)^{-14}$, I = 114, I/N - 0.115

In each section of the table the first line shows f = the number of individuals who were counted as belonging in the double classification indicated by the vertical and horizontal headings. Thus the entry in the top line of the left hand column shows (AB) = 274, which is the counted number of individuals who were predicted to pass and who also passed.

In the second line of each section of the table is the entry f_o = the number of individuals who would have been included in the double classification if the prediction had been independent of the outcome. In the present instance (B) = 556, the number of individuals who passed the course, while 990 is the number of individuals who either passed or failed. If the basis of prediction had been independent of the outcome, then out of any sample that was predicted to pass, the number who passed should have been proportional to 556/990 - 56.16 per cent. Under this condition, from any group of students who were predicted to pass, 56.16 per cent should most probably pass and 43.84 per cent should most probably fail; likewise, from any group of students who were predicted to fail, 56.16 per cent should pass and 43.84 per cent should fail. These numbers predicted by the independency hypothesis are set down in parentheses in the second line of each block of the table, and are labeled "f_o."

In the third line of each block is shown the difference $\Delta = f - f_o$ between the number of individuals counted in each of the doubly defined classes and the number which the independency hypothesis predicts. The definition of $\Delta = (AB) - (A)(B)/N$, in which (AB) is the number who were predicted to pass and also passed, (A) = the number predicted to pass, (B) = the number who passed, and N = the total number of students in the sample. If the sign of Δ is positive, it means that the number of successful predictions exceeded the number of unsuccessful predictions. In fact, $2\Delta = (AB) - (A)(B)/N + (A'B') - (A')(B')/N = 57 + 57 = 114$ is the excess of correct predictions over the number of correct predictions which would occur by mere chance. We define $\Gamma = 2\Delta$ as the absolute gain in the number of correct predictions which we get by using the test instead of guessing, and Γ/N as the relative gain. Thus, since $\Delta = 57$, $\Gamma = 114$, while $\Gamma/N = 114/990 = 0.1152 - 11.52$ per cent. In other words, 114 more students, constituting 11.52 per cent of the total number, were correctly classified by the test who would have been incorrectly classified by mere chance.

If the sign of Δ had been negative, so would have been the sign of Γ and of Γ/N. This would have meant that the test ran counter to outcome, and that one had better bet against the prediction than with it. However, once the rule is established, we could use a negatively predicting test for positive prediction, just as we could use a backward running clock as well as a forward running clock for telling the correct time. All that we need to do is reverse the order of the scale numbers from an arbitrary zero point.

The fourth line in each section of the table denotes Δ^2. Since in any 2 x 2 table every value of Δ is the same, so is every value of Δ^2.

The fifth line in each section of the table denotes $x^2_o = \Delta^2/f_o$, i.e., the square of the discrepancy divided by the theoretical frequency. For any single cell, this entry may be considered as x^2/σ^2 in a table of the normal probability integral, from which one derives the probability of so large a discrepancy occurring by chance. For example, in the (AB) class, we have $f = 274$, $f_o = 217$, $\Delta = 57$, $\Delta^2 = 3249$, $x^2_o = \Delta^2/f_o = 14.97$, $x_o = 3.87$. From the probability integral table we derive that the probability P of getting so large a positive deviation from zero as $x_o = 3.87$, is of the order of $5(10)^{-5}$. This alone is enough to show that the outcome is not independent of prediction. If we sum the values of x^2_o -- and this we can do because x^2 is additive -- for the individual cells, we get $x^2 = 55.97$. Either by calculation or by the use of such tables as Table IV in Pearson's Tables for Statisticians and Biometricians, Part I, third edition, we derive that the probability P of the discrepancies being jointly due to chance is of the order of $3(10)^{-14}$. Thus the independency hypothesis is contradicted; there is some causal relationship between prediction and outcome although it remains to determine what it is.

If the examiner who makes the prediction knows something about the student's progress, as well as the facts brought out in the interview, it is very hard to prevent his prediction from being swayed by this additional knowledge. This bias may occur without the least mal-intention on the part of the examiner, especially if he is clinically minded, and therefore inclined to base his judgment on all his knowledge of the individual considered "as a whole." But it is just this inclination which makes it hard to classify clinical judgments rigidly, as one must do if one can treat them justly by statistical procedures. There are reasons, not apparent in Table 3 or in any of the similar tables from which it was compiled, for wondering if some of these predictions were not influenced by knowledge of other facts than those which the interview would normally bring out.

An over-all control experiment

Certain other medical officers at the station therefore made a new experiment, which required the interviewer to make all his predictions in ignorance of any training record that the students may have made. This work was done by Lieut. Cmdr. Rex H. White (MC) USNR under very stringent precautions. Dr. White made over-all ratings on 182 students in classes 73-78. Of these students, 139 completed the course and 43 were dropped. Dr. White divided those students into five classes, according to the over-all favorability of prediction. We shall designate them arbitrarily, showing first his symbols and then our own: AA, V; A , W; A, X; A-, Y; BA, Z. The class denoted BA, Z was a class which he judged to be "below average," in respect to some ideal norm.

The results in so far as they concern passing versus failing may be summarized in Table 4:

TABLE 4

EXAMINER'S OVER-ALL RATINGS IN RELATION TO PROBABILITY OF PASSING

		Over-all ratings					
		V	W	X	Y	Z	Total
Passers	f	21	20	30	29	39	139
	f_o	17	17	30	26	49	139
	Δ	4	3	0	3	-10	0
	Δ^2	16	9	0	9	100	-
	x^2_o	.94	.53	0	.35	2.08	-
Failers	f	1	2	10	5	25	43
	f_o	5	5	10	8	15	43
	Δ	-4	-3	0	-3	10	0
	Δ^2	16	9	0	9	100	-
	x^2_o	3.20	1.80	0	1.13	6.67	-
Total	f	22	22	40	34	64	182
	f_o	22	22	40	34	64	182
	Δ	0	0	0	0	0	0

$x^2 = 16.67$, $P = 2(10)^{-3}$

It is quite obvious that there is an association between these ratings and the probability of completing the course, in the sense that those who received more favorable ratings than Z contained the larger proportion of completers. On the other hand, as may be seen from the third line or the eighth line of the table, the test displaced only 20 applicants from the classification which they would have had by chance. It is interesting that the distribution of passers and failers in class X is exactly what chance would have yielded. It is possible, however, to summarize these results more usefully in other forms: namely, Tables 5 and 6. Table 5 shows what would happen if the Navy had accepted classes V, W, X, Y, and rejected class Z.

TABLE 5

		V,W,X,Y	Z	Total
Passers	f =	100	39	139
	f_o =	90	49	139
	Δ =	10	-10	0
	Δ^2 =	100	100	-
	x^2_o =	1.00	2.04	-
Failers	f =	18	25	43
	f_o =	28	15	43
	[illegible] =	-10	10	0
	[illegible] =	100	100	-
	x^2_o =	3.57	6.67	-
Total	f =	118	64	182
	f_o =	118	64	182
	Δ =	0	0	0

$x^2 = 13.28$, $P = (10)^{-4}$, $\Gamma = 20$, $\Gamma/N = 0.110$

Since the probability P of the association between prediction and outcome being due to chance is of the order of $(10)^{-4}$, i.e., of 1/10,000, we have to regard the correspondence as genuine. Since the examiner's ratings were made without independent knowledge of the students' progress, other causes of the association need to be sought for carefully. The meaning of $\Gamma = 20$, $\Gamma/N = 0.11$ is that the sample of 182 students contained 20 individuals, 11 per cent of the total number, who would have been improperly classified by chance but were properly reclassified as passers or failers by Dr. White's over-all ratings. In other words, if the Navy had prescribed these ratings as additional qualifications for candidates who had passed its other selection tests, it would have excluded 64 candidates, of whom 25 were failers. These 25 individuals constitute 58 per cent of the 43 failers. To get rid of them, the Navy would have lost the services of 39 passers also. These 39 individuals make up 28 per cent of the 139 passers. Had the Navy employed these standards of selection, it would have retained 118 applicants, of whom 100 individuals, 85 per cent, were passers. This ratio is to be compared with 139/182 = 76 per cent of passers in the original sample. If it could replace the 64 individuals rejected by the test by 64 persons who qualified for classes V, W, X, Y, of whom most probably 85 per cent would have been passers, it would again have 182 applicants, of whom 154 individuals instead of 139 were passers, and 28 individuals instead of 43 were failers. One can conceive some circumstances in which this move would be desirable.

However, it is interesting to see what would happen if we assemble into one class the V, W, raters, expecting them to pass, and into another class the X, Y, Z raters, expecting them to fail. The results are shown in Table 6.

TABLE 6

		V,W	X,Y,Z	Total
Passers	f =	41	98	139
	f_o =	34	105	139
	Δ =	7	-7	0
	Δ^2 =	49	49	-
	x^2_o =	1.44	.47	-
Failers	f =	3	40	43
	f_o =	10	33	[illegible]
	Δ =	-7	7	[illegible]
	Δ^2 =	49	49	[illegible]
	x^2_o =	4.90	1.48	-
Total	f =	44	138	182
	f_o =	44	138	182
	Δ =	0	0	0

$x^2 = 8.29$, $P = 2(10)^{-3}$, $F = 14$, $F/N = 0.077$

As shown in the second column of the table, these ratings would reject 138 applicants out of 182, or about 76 per cent. Of those rejected, 98 individuals, about 71 per cent of those who were rejected, would have been passers while 40 individuals, 29 per cent of those who were rejected, would have been failers. But of the 44 applicants who were retained, 41 individuals, making up 93 per cent of those who were retained, would have been passers, while 3 individuals, making up about 7 per cent of those retained, would have been failers. Could the Navy have afforded to reject 76 per cent of the applicants who had passed the preceding selection devices, in order to reduce its proportion of failers from 24 per cent to 7 per cent? It could, provided it had a reservoir from which it could draw 580 more applicants like these. The 24 per cent who could have qualified for classes V, W would have provided the 138 individuals whom it rejected from this sample of 182; and of the 182 candidates whom it had thus accepted, 169 individuals making up 93 per cent of the total, would most probably have been passers. The probability that the discrepancies are jointly due to chance is still about 1 in 500, and therefore probably negligible.

Our account of Dr. White's procedures and results is accurate but incomplete. He is the proper person to describe in detail his procedures. We have summarized only some of the most important results, and have considered only some of the possible interpretations, not because the experiment was lacking in importance, but because both he and the other interested officers at the station believed that another mode of attack might be still more fruitful than this.

Controlled interviews

Their hypothesis, briefly stated, was that it might be more useful to select certain personality traits, or items of biographical information, and attempt to determine them as nearly objectively as possible, than to make over-all ratings. Therefore they picked out a good many, and considered them separately. The results appear in Table 7, but before we can usefully scan the table, we should examine definitions of traits.

In respect to the trait "Inheritance," I quote from a letter of Comdr. Victor S. Armstrong (MC) USN, which shows that the examiner's judgment is based on an alternation of certain sets of conditions. These are as follows:

Good: No history of psychosis, psychoneurosis, epilepsy, alcoholism, etc., in two generations of forebears.

Fair: No history of major psychosis or epilepsy in forebears. May be emotional upsets such as quick temper, a mild nervous breakdown in one parent, poor adjustments on part of one parent, etc.

Poor: History of definite psychosis, epilepsy, etc., in one or both parents, or alcoholism in immediate parents.

The trait called "Environment" is similarly defined, as an alternation of simultaneous conditions, described as follows:

Good: Family circumstances comfortable and agreeable. Community life good either rural or urban.

Fair: Ignorance and straitened circumstances in home. Poor emotional responses on part of one or both parents. Recreational and educational opportunities curtailed.

Poor: Orphanage life - unreasonable or alcoholic parents. Bickering and constant unpleasantness in home. Police record on part of parent, brother, or sister. Unhappy life. Divorce, etc.

Similarly, the most important characteristic of all, which the examiners called "Desire to fly" is evaluated according to some such set of rules as the following:

Strong: Sustained desire over considerable period of time and objective manifestation of interest such as building model airplanes, taking rides at own expense or inconvenience. Entry into service or pursuit of education with flying as goal idea.

Medium: Likes the experience, the people, and the additional pay.

Poor: Taken as an escape from unsatisfactory situation. To get out of an unpleasant fate or just a chance to make a living. No basic desire.

The trait called "Education" is defined by the kind of school in which the candidate's formal training was terminated; namely, grade school (G), high school (H), naval academy (A), college or technological school of college grade (C), or university work after graduation (U). The candidates' own statements, backed up by their transcripts were used. To assign a student to any of these classes above G, it was necessary that he complete at least half the training prescribed for this class.

The classes "Version" and "Ambiversion" are derived from Bernreuter's test scores, "Version" being broken down into "Ambiversion," "Introversion," "Extroversion," and the class "Ambiverts" being distinguished from the class of "non-Ambiverts," in the sense that the latter either did not qualify as "ambiverts" or else were not tested for that trait. Since the predictive value of this classification is of the order of 0.01, we need not consider these classifications farther than to note that they have been made and diligently evaluated.

The trait "Emotional stability" is derived from the unanalyzed impression which the subject made on the interviewer; it is variously denoted as "Good" (G), "Undetermined" (U), "Poor" (P). In only one group of 71 students were unequivocal judgments rendered. The result suggests that it might have paid to render this judgment on every applicant.

The trait called "Aggression" likewise depends on unanalyzed impressions made by the subject on the interviewer. It is rated "Good" (G), "Undetermined" (U), "Poor" (P), on 519 students, but as either G or P on only 148 students. However, since its predictive value is nil, we need not consider it farther now.

The trait called "Self-confidence" also was evaluated subjectively by the interviewer, and rated as "High" (H), "Undetermined" (U), or "Low" (L). It was undetermined in 434 instances out of 595, and determined in 85 instances. Among the latter it yielded a Tschuprow coefficient of contingency $T = 0.14$, and a predictive value of 0.01. These terms will be defined presently; meanwhile we may say that this judgment taken alone is practically worthless, but if it had been rendered on all subjects, and had maintained this predictive value, it might have been usefully included in a battery of similar tests, on the principle that "everything added to what you've got makes just a little bit more."

The trait called "Responsibility" was evaluated for only 43 subjects, and by its predictive value, which is about 0.07, it showed some promise of usefulness if it were made part of a test battery. Its evaluation depended on answers to such questions as the subjects' acceptance of regular chores, such as the care of domestic animals, the lawn, or his endeavors to earn spending money, money for clothes, or money needed to help support the family. Despite this poor showing, the face validity of this item seemed to warrant its inclusion in biographical inventories which were later developed with this inventory as a guide.

We need not describe in detail the manner in which the interviewer's judgments of "Frankness, Tension, Reactian time, Over-cautiousness, or Courage" were determined since none of them turned out to have an appreciable predictive value. All were based on the interviewer's impressions acquired during the interview.

These results are summarized in Table 7. The second column of this table gives the source of this writer's information, as follows: (1) A report from Commander Armstrong to the Medical Officer of the U. S. Naval Air Station, Pensacola, dated 29 December 1939; (2) a report by Dr. Armstrong to the same authority, dated 22 April 1940; (3) an undated memorandum by Dr. Kellum covering psychological research between 2 January and 15 February 1940. Through the kindness of these officers and the Medical Officer, Captain Frederick Ceres (MC) USN, I was allowed to digest their findings. The results of my study were duly reported to this Committee in the autumn of 1941 in a memorandum entitled "Personal data as predictors of aeronautical success: a note on certain information collected by medical officers at the U. S. Naval Air Station, Pensacola, Florida."

One copy of this report was sent to the Medical Officer of the station, and another was made for the Medical Research Division, Bureau of Aeronautics, U. S. N.

The third column of Table 7 shows the basis of classification of the interviewer's ratings. As illustrated by the first three lines, for example, the judgments of "Inheritance" were categorized as G, F, P, U, i.e., "good," "fair," "poor," and "undetermined."

The first line shows that from source A only three classes were used and considered separately, namely G, F, P, the class U being omitted from consideration. The second line shows that from source B only classes G and U were considered. The third line shows that from sources A and B combined, those belonging to class G were considered together, and those belonging to classes F, P, or U were considered together. In general, inclusion of class symbols in parentheses means that the included classes were consolidated for comparison.

The constant Chi-square = $\Sigma(f - f_0)^2/f_0$ is derived by the procedure described above and is illustrated in Tables 3, 4, and 5.

In the seventh column of Table 7, one finds Tschuprow's coefficient of contingency T.[1] [Its definition and properties are described in Chapter 5, section 12, of Yule and Kendall's "Introduction to the Theory of Statistics," 11th edition, London, Griffin, 1937.] It is shown solely for the information of statistically minded readers, who are used to interpreting contingencies in accordance with a parameter called the Pearsonian coefficient of correlation r.[1] [In contingency tables of this kind, T is strictly analogous to r and in certain limiting instances is equivalent to r.] The non-statistical readers may disregard it in the remainder of this discussion, except to note that the so-called predictive value E of the test is based on T in this manner: namely, $E = 1 - \sqrt{1 - T^2}$. In a word E indicates that proportion of the standard error of estimation based on guessing which one removes by using the test.

Table 7 shows that several individual items have a predictive value which is high enough to justify their inclusion in a test battery. Chief among them is the so-called "Desire to fly," which in every classification in which it is used serves to differentiate the passers from the failers to a useful degree. If the information had been available, our next move would have been to resort to multiple correlation or multiple contingency, but unfortunately, the records were not so kept as to show every item of information for every person, and the cost of running them down would have exceeded the funds at the disposal of this experimenter. This desideratum was attained, however, in some classes which the officers of the station studied later.

Bernreuter's personality tests

In 1931 and 1935, Bernreuter put out an inventory of 125 items which he called a personality test. The inventory comprises 125 questions most of which are ambiguous. Some are ambiguous because they imply unstated dates or epochs. Some call for censuses that the subject cannot have made, and which are implied in such words as "usually," "more than," etc. Some call for evaluations that the subject cannot make hastily, and perhaps not at all. They are implied in such undefined terms as "frequently," "often," and the like. All such questions are ambiguous, and intended to be so. However, categorical answers are demanded. Explanations of such questions are the following:

Do you day dream frequently?

Do you prefer to associate with people who are younger than yourself?

Do you consider yourself a rather nervous person?

Are you inclined to study the motives of other people carefully?

TABLE 7

ASSOCIATION BETWEEN EXAMINERS' JUDGMENTS OF CERTAIN PERSONALITY TRAITS AND SUCCESS OR FAILURE IN NAVAL AVIATION TRAINING

Test	Source of information	Basis of classifi-cation	Number of students	Value of Chi-square	Probability that success is independ-ent of trait	Coefficient of contin-gency T	Predictive value of test
Inheritance	A	G, F, P	513	1.63	0.46	0.05	0.00
Inheritance	B	G, U	439	9.20	$(10)^{-3}$	0.15	0.01
Inheritance	A+B	G, (F, P, U)	952	6.65	$3(10)^{-3}$	0.08	0.00
Environment	A	G, (F, P)	520	8.08	0.02	0.13	0.01
Environment	B	G, U	439	15.20	$(10)^{-5}$	0.19	0.02
Environment	A+B	G, (F, P)	959	14.63	$7(10)^{-5}$	0.12	0.01
Desire to fly	A	S, M, I	519	64.92	$8(10)^{-15}$	0.30	0.05
Desire to fly	B	S, M, I	439	107.37	$(10)^{-23}$	0.42	0.10
Desire to fly	B	S, (M, I)	439	87.86	$(10)^{-21}$	0.45	0.11
Desire to fly	A+B	S, (M, I)	966	139.48	$(10)^{-32}$	0.38	0.08
Education	A	G, H, A, C	516	40.31	$(10)^{-9}$	0.21	0.02
Education	B	G, H, A, C	445	48.76	$(10)^{-10}$	0.25	0.03
Education	B	H, (A, C)	375	3.53	0.03	0.03	0.00
Education	A+B	G, H, A, C	961	77.86	$(10)^{-17}$	0.22	0.02
Education	A+B	G (H, A, C)	961	55.15	$(10)^{-14}$	0.24	0.03
Education	A+B	H, A, C	796	23.53	$7(10)^{-6}$	0.14	0.01
Education	A+B	H, (A, C)	796	15.75	$4(10)^{-5}$	0.14	0.01
Education	A+B	H, A	508	2.58	0.05	0.06	0.00
Education	A+B	A, C	502	11.85	$3(10)^{-4}$	0.15	0.01
Education	B	(C, A), U	439	20.13	$5(10)^{-6}$	0.21	0.02
Education	B	C, U	439	11.91	$3(10)^{-4}$	0.21	0.02
Version	B	A, I, E	440	6.27	0.01	0.10	0.01
Ambiversion	B	A, U	447	6.08	$7(10)^{-3}$	0.12	0.01
Emotional status	A	N, I, E	490	1.90	0.40	0.05	0.00
Emotional control	A	G, F, P, U	519	6.53	0.09	0.09	0.00
Stability	A	G, U, P	519	5.33	0.04	0.05	0.00
Stability	A	G, P	71	6.48	$5(10)^{-3}$	0.30	0.05
Agression	A	G, U, P	519	1.06	0.50	0.04	0.00
Agression	A	G, P	148	1.02	0.16	-0.08	0.00
Attention	A	G, U, L	519	1.53	0.50	0.05	0.00
Attention	A	G, L	92	2.73	0.05	0.17	0.02
Attentiveness	B	G, F, P	231	0.05	0.50	0.04	0.00
Self-confidence	A	H, U, L	519	0.74	0.40	0.03	0.00
Self-confidence	A	H, L	85	1.81	0.22	0.14	0.01
Responsibility	A	G, U, L	519	7.09	0.03	0.10	0.01
Responsibility	A	G, L	43	5.54	$(10)^{-3}$	0.36	0.07
Frankness	A	F, U	519	0.95	0.67	-0.04	0.00
Tension	A	T, U	519	3.87	0.02	0.09	0.00
Immaturity	A	I, U	519	0.005	0.48	0.00	0.00
Reaction time	A	G, S	142	0.16	0.35	0.03	0.00
Over-cautiousness	A	O, U	519	1.49	0.11	0.05	0.00
Courage	A	G, U, L	519	5.91	0.05	0.09	0.00
Courage	A	G, (Q, P)	137	4.76	0.02	0.19	0.02
Marital status	B	M, S	436	0.002	0.52	0.00	0.00

Have you ever had spells of dizziness?

Do you usually prefer to work with others?

To each question the subject can answer yes (Y), no (N), doubtful (?), or omit the answer (O). An answer Y or N to most of these questions is logically meaningless, but it would be rash to presuppose that the subjects' answers are determined solely by factual and logical considerations.

Bernreuter formulated some scoring devices which accompany the test. Some of them give estimates of what he calls "neurotic tendencies," B1-N; others of what some call "crippling self-sufficiency," B2-S; still others of what he calls "introversion," B3-I; "dominance," B4-D; or "self-confidence," F1-C. This writer took the data presented in Dr. Kellum's report, and subjected each to the treatment indicated below.

In respect to the trait called "neurotic tendencies," B1-N, and "crippling self-sufficiency," B2-S, Pearson's Chi-square test for homogeneity as between passers and failers is satisfied within the limit P - 0.70 and 0.80, respectively. Since this test is more sensitive than the test of reliability of difference between means, I did not apply the latter.

In respect to the trait called "introversion," B3-I, the results may be summarized in Table 8. The criterion classes are as follows:

5.	Passed without extra time	9	individuals
6.	Passed with not more than 9 hours extra time	23	"
7.	Passed with more than 9 hours extra time	4	"
8.	Passed after being turned back or repeating part of the course	9	"
9.	Failed	25	"

In Tables 8, 9, and 10, each difference between mean scores on the Bernreuter test is taken with the mean score belonging to the class denoted by the columnar heading as minuend; and the class denoted by row heading as subtrahend. The upper number denotes the difference in this sense, the lower number denotes the probability of so large a directional difference occurring by chance.

Table 8 shows that the score for introversion does not distinguish any of these classes from any other within the standard P = 0.01. The difference between the mean score for class 7 (passed with more than 9 hours of extra time) and class 9 (failed) looks interesting but since class 7 included only

TABLE 8

Introversion - B3-I

Difference between mean scores (upper entry) and probability of the difference occurring by chance (lower entry) between members of the various progress classes.

Criterion classes	Criterion classes 5	6	7	8
6	-15.43 .93	-	-	-
7	19.44 .73	34.87 .08	-	-
8	-16.67 .73	-1.24 .07	-36.11 .10	-
9	-21.48 .09	-6.05 .32	-40.92 .05	-4.81 .39

4 students, one had better not trust appearances. This class, however, does tend by every test to distinguish itself from the others, except class 5 -- almost within reliability standards.

This writer caused the returns for each subject to be scored for "dominance," B4-D. The results are summarized in Table 9.

TABLE 9

Dominance - B4-D

Difference between mean scores (upper entry) and probability of the difference occurring by chance (lower entry) between members of the various progress classes.

Criterion classes	Criterion classes 5	6	7	8
6	9.28 0.30	-	-	-
7	-34.64 .30	-43.92 .04	-	-
8	17.89 .60	8.61 .38	52.53 .06	-
9	7.19 .34	-2.09 .14	41.83 .04	-20.70 .35

Table 9 shows that the score for dominance, B4-D, does not distinguish any class of students from any other, within the limits which are usually accepted as reliability standards. A near exception is class 7, which we have mentioned before.

The writer also had the records scored for "self-confidence," F1-C. The results appear in Table 10.

TABLE 10

Confidence - F1-C

Difference between mean scores (upper entry) and probability of the difference occurring by chance (lower entry) between members of the various progress classes.

	Criterion classes			
Criterion classes	5	6	7	8
6	-19.91 .25	-	-	-
7	64.19 .07	84.10 .02	-	-
8	-23.23 .29	-3.32 .47	-87.42 .05	-
9	-13.04 .32	6.87 .39	-77.23 .03	10.29 .40

The test scores do not distinguish any criterion class from any other within the usual limit P - 0.01. The differences between the means for class 7 and classes 5, 6, 8, and 9 would be suggestive if class 7 contained many members. All probability rules become doubtful when the class population is as low as 4.

The report of Dr. Kellum covering the period January 2 - February 15, 1940, contains a comparison of the scores of 26 passers and 31 failers on Bernreuter's test B1-N, "neurotic tendencies." The data were made available in a form which permitted the writer to apply Pearson's test for homogeneity. The results showed that passers and failers can be regarded by Pearson's standard as having been drawn from the same parent population by the same sampling procedure, the probability of the discrepancies being jointly due to chance being of the order of 0.70.

The same report presents a comparison of the scores on Bernreuter's test for "crippling self-sufficiency," B2-S, of 26 passers and 32 failers. Again the results showed that the two groups can be regarded as having been drawn from the same parent population by the same sampling procedure, the probability of the discrepencies being jointly due to chance exceeding 0.80.

Here the present writer tenders some remarks of his own. Bernreuter's test, like Strong's interest blank, Woodworth and Wells' neurotic inventory, and many other personality tests, requires the subject to return categorical answers to ambiguous questions.

Such a demand is bound to offend a subject if he considers the question logically, for he knows that since the question is logically meaningless, any answer that he may give may mislead the reader. As E. T. Bell remarks[2] if anyone refuses to answer such a question in the terms in which it is stated, his refusal counts for a mistake; if he returns any answer in such terms, his answer is wrong. There is no provision in the rating scheme for crediting any intelligence to an individual who refuses to participate in a nonsensical game. From the fact, however, that almost all these cadets, and nearly all the officers who are subjected to these questionnaires do return categorical answers to most of the questions, we have to treat their answers as non-logical, as being returned in a spirit of resolution to conform if possible, and as expressing some vague emotional attitudes which are hard if not impossible to describe. The fact that the same subjects return fairly consistent answers at widely separated dates is most interesting but it proves nothing except that the test situation seems to give rise to attitudinal habits, which tend to be reinstated when the test situation is reproduced. Just as a tendency toward an error in perception or judgment may become habitual, as well as tendency toward accurate judgments, so do those tendencies become habitual also. It is a fact that in the statistical sense, these questionnaire returns have a high repetitive reliability. It is disturbing that nobody seems to know how to interpret it.

Dr. Kellum and his colleagues kept records of the flight grades of 70 subjects, together with records of their scores on B3-I (introversion), B4-D (dominance), and F1-C (self-confidence). One of my assistants, Gloria Ladieu, reported the data in Table XI.

The only differentiation which is noticeable lies in the score on B4-D which is correlated -0.16 with flight grade. This coefficient satisfied a standard of reliability that [illegible] some investigators though not others. However, it runs in the opposite sense from Bernreuter's expectation. One reason for the low correlation may lie in the character and the spread of the flight-grades themselves. Although the average grades are expressed by three-place numbers, nevertheless, these numbers are averages of numbers which are first of all nominal, not cardinal, and which are admitted to be inaccurate in the second digit. The standard

[2] Bell, E. T. Numerology, Baltimore, Williams and Wilkins Company, 1934, P.166.

TABLE 11

Correlation between flight grades of 70 subjects and their scores on Bernreuter's tests indicated below

Test	Mean score	Standard deviation	Correlation with flight grade	Predictive value of test
B3-I (Introversion)	-57.8	44.0	0.04	0.00
B4-D (Dominance)	74.2	52.9	-0.18	0.02
F1-C (Self-confidence)	-75.0	84.6	0.13	0.01

deviation is of the order of 0.132 which is about 4.6 per cent of the average. Hence, honor students, ordinary students, and failing students tend to receive nearly the same average flight-grade. However, since Bernreuter's test scores do not correlate appreciably with any other criterion of final performance or economy in learning, perhaps we need not worry about their feeble and uncertain correlation with a variable which is so nearly constant as average flight-grade.

Item analysis of Bernreuter's test

We have seen that Bernreuter's personality inventory did not distinguish passers from failers, or various classes of passers from each other, or students having different averages of flight grades. (The last mentioned failure, however, is probably due to the fact that the flight grades vary but little from one student to another, and do not in themselves distinguish passers from failers at all well.) But certain other evidence, which I shall not discuss in this report, convinced me that whether Bernreuter's inventory is scored according to Bernreuter's rules or Flanagan's, no trait has yet been arbitrarily defined by Bernreuter's test scores that will be usefully correlated with any such variables as final success, extra time, and other indices of extra cost in training, progress in a given stage, and the like. The interested reader may test the factual truth of this conviction for himself. At the present time, properly authorized persons can get the raw data for compilation by applying through proper channels.

Since the usual procedure seemed to be unpromising, I decided to correlate each of the 125 items with the pass or fail criterion; to identify and count those items which have useful predictive value; to determine by using Pearson's x^2 test of goodness of fit to the ideal distribution whether the number of useful items exceed the number demanded by chance; and to see whether some mode of combining the valuable items could be found to maximize their predictive value as a battery or set.

In considering each item, the first step was to count separately the passers and failers who recorded "Yes," "No," "Doubtful," or who "Omitted answer." The next step was to calculate the corresponding numbers demanded by the independency hypothesis, using the method described in Yule and Kendall's "Introduction to the theory of statistics," 11th edition, section 22.16, and obtain the value of x^2 and the corresponding probability P of the discrepancies between census and expectation being due to chance. The results were reported to the Committee 20 December 1940, and a supplementary report was sent in 7 January 1941. In a word, we found by count nine items which yielded values of x^2 as great as 9.24 or greater, whereas the independency hypothesis allows fewer than one item.[3] The probability is negligibly small that the excess of 9 useful items is due to chance. At least some of them, and quite probably all, are genuine.

The nine apparently useful items are as listed below along with the obtained values of x^2, P, T, and E. Here T denotes Tschuprow's coefficient of contingency, which is the closest approximation to Pearson's coefficient of correlation r between type of answer and success or failure that can be made in the circumstances; while $E = 1 - \sqrt{1 - T^2}$ may be regarded as the coefficient of effectiveness of the item in separating passers from failers.

Item 11: Do you try to get your own way even if you have to fight for it? $x^2 = 11.86$, $P = 3(10)^{-3}$, $T = 0.14$, $E = 0.01$.

Item 18: Are you "touchy" on various subjects? $x^2 = 9.45$, $P = 9(10)^{-2}$, $T = 0.12$, $E = 0.01$.

Item 44: Have books been more entertaining to you than companions? $x^2 = 14.50$, $P = 7(10)^{-4}$, $T = 0.15$, $E = 0.01$.

Item 50: Do you usually try to avoid arguments? $x^2 = 9.87$, $P = 7(10)^{-3}$, $T = 0.13$, $E = 0.01$.

Item 63: If you are dining out do you prefer to have some one else order dinner for you? $x^2 = 9.92$, $P = 7(10)^{-3}$, $T = 0.13$, $E = 0.01$.

Item 104: Do your feelings alternate between happiness and sadness without apparent reason? $x^2 = 9.25$, $P = 0.01$, $T = 0.12$, $E = 0.01$.

Item 113: If you are hiking with a group of people, where none of you knew the way, would you probably let some one else take the full responsibility for guiding the party? $x^2 = 14.11$, $P = 9(10)^{-4}$, $T = 0.15$, $E = 0.01$.

[3] This value of 9.24 taken with two degrees of freedom, corresponds to a probability $P = 0.01$ in favor of the discrepancies being jointly due to chance. We did not feel justified in putting into a test battery any item for which P exceeded 0.01, although some statisticians might have done so.

Item 115: Are you often in a state of excitement? $x^2 = 10.42$, $P = 5(10)^{-3}$, $T = 0.17$, $E = 0.015$.

Item 122: Can you be optimistic when others about you are depressed? $x^2 = 10.28$, $P = 6(10)^{-3}$, $T = 0.13$, $E = 0.01$.

None of these items is important for its predictive value when it is taken alone. Nevertheless, if we weight them in the best possible manner, they yield an unshrunken and hardly reliable multiple correlation of 0.35 with the pass-fail criterion. This combination of Bernreuter's items might therefore be useful if it were put into a battery with items of other kinds, taken from other tests. However, a better way of handling them was available, which we shall describe in Part III of this report.

Meanwhile, I recommended to the Committee that these items or their equivalents be included in some new biographical inventories which the Committee was considering for development. In making this suggestion, however, I mentioned two reasons why one should restrain one's expectations of the results. The first reason is the well known fact that in the original application of this test, each of the valuable questions is presented in a context. It is quite probable, although the fact is still uncertain, that the context partly determines the subject's answers. If the question is presented in a new context, its predictive value may be either increased or diminished. Quite often, it is diminished.

The second reason for advising caution is the manner in which the variance between the actual number and the expected number of various types of answers to a given question is distributed among the various categories of response. In the report dated 20 December 1940, I exhibited the evidence in detail. For example, in respect to Item 11, quoted above, 63 subjects out of 429 answered the question "?." Of the 63 doubters, 29 instead of 41 were passers, and 34 instead of 22 were failers. Hence, there is a deficiency of 12 passers in this category and an excess of 12 failers. The value of x^2 in these two cells is 3.51 and 6.55, respectively, making a total of 10.06, or about 85 per cent of the total variance within the entire distribution. In fact all the remainder of the total variance can be attributed to chance. Thus this answer served to differentiate 12 persons out of 429, i.e., 2.8 per cent of all the subjects, and half of all whom the item differentiated. Although the usual probability formulas imply that this result is statistically dependable, we could profitably review the question whether such a situation as this is wholly included within the scope of the usual laws of probability. We cannot discuss the question in detail here for it belongs to theory. However, it may not have been finally settled.

PART III: THE MOST EFFECTIVE COMBINATION OF BIOGRAPHICAL INFORMATION

The answers to Bernreuter's questionnaire could be regarded as biographical information in the sense that they purport to indicate some of the individual's likes and dislikes, habits of social adjustment, interests, etc. It is true, of course, that since many of the questions are meaningless or indefinite, so must be the answers if one considers them from a logical and factual point of view. For example, consider question 115: "Are you often in a state of excitement?" The adverb "often" is ambiguous, for what one person might call often, another might or might not; while for the same person, one event that recurred, say, 14 times a week might seem to recur often while another event which recurred with the same frequency might seem to recur seldom. It is a cardinal principle of counting that only those things or events should be counted together which are of the same kind: i.e., those things or events which qualify for inclusion in the same class. In this instance we are not justified in counting together two or more typical answers to such a question as this unless we disregard their meaning to the subject and consider only their grammatical similarity. This, in fact, is what the inventors of such tests as these instruct us to do. And nothing in the answers which the subject returns will tell us whether he is trying to describe himself as he believes himself to be, or as he would like to be or as he would like to appear to an onlooker. We know, of course, that his intent helps determine his answers.

However, we can take the answers to the differentiating questions as primitive facts, and combine them in the most effective manner with other items of biographical information, some of which are more objective than this.

Such information was collected in whole or in part on the following classes: 80-O, 81-C, 82-C, 83-C, 86-C, 87-C, 88-C, 89-C, and 90-E. The terminal letters O, C, and E, denote the make-up of the class, as Officers, Cadets, or Enlisted men, respectively. The several variables considered were compiled from the records of the station by Mary L. Boots, Research Associate, and Oscar Backstrom, Jr., Research Assistant, under the immediate direction of Mrs. Boots.[4] The variables used were as follows:

05	P/P'	Ultimate success or failure in the course.
06)	TT	Total time spent in training. Excess passenger and instrument unit time not included.
07)		

[4] I must say here that this part of the work must be done impeccably or all which follows it is infected with uncertainty. The records of every flight had to be carefully studied, and supplemented by records from at least five other sources. Some of the original records were self-contradictory and several classes of data had to be discarded for this reason. The classes of data that were retained are unambiguous and dependable. This phase of the program cannot be entrusted to novices or mere clerks.

08	TET	Total extra time allowed during course of training.
09	TECT	Total extra check time (spent in taking extra checks).
10	TUDT	Total used dual time in course.
11	TEC	Total number of extra checks in course.
12	TA	Total accidents in course.
13	TV	Total violations in course.
14	TNA	Total near-accidents in course.
15	Stage	Stage in which dropped for failers in course, or if not dropped with uncompleted stage, last completed stage coded. Passers all the same.
16) 17) 18)	Peckham	Peckham score devised by HMJ and RHP, according to number of "up" or "down" checks.
19	Navig.	Number of times navigation course was failed in ground school.
20	GSC	Number of ground school courses failed.
21	Age	Student's age at beginning of course.
22	Educ.	Education of student at time of entering.
23	M.Int.	Major interest or field in college.
24	Pr.M.Exp.	Previous military experience.
25	Ath.	Athletic participation.
26	Relig.	Religion.
27	Des.Fly	Desire to fly.
28	Hered.	Heredity.
29	Env.	Environment.
30	Maj.Oc.	Most recent prior major gainful occupations.
31	Empl.Sch.	Employment during school years.
32	Pr.A.Tr.	Previous training in aeronautics.
33	Fitness	Fitness, examiner's judgment of.
34	Bern.11	Bernreuter Question No. 11.
35	Bern.18	Bernreuter Question No. 18.
36	Bern.44	Bernreuter Question No. 44.
37	Bern.50	Bernreuter Question No. 50.
38	Bern.63	Bernreuter Question No. 63.
39	Bern.104	Bernreuter Question No. 104.
40	Bern.113	Bernreuter Question No. 113.
41	Bern.115	Bernreuter Question No. 115.
42	Bern.122	Bernreuter Question No. 122.
43	Bern.B1N	Bernreuter B1N difference score.
44	Bern.B2S	Bernreuter B2S difference score.
45	Bern.B4D	Bernreuter B4D difference score.
46	Morgan	Morgan scores.
47 thru 79	ET 1-A	Extra time, Squadron I, Stage A, and so on for every stage of every squadron to total amount of extra time for Squadron V.
80 thru 94	ECT 1-A	Extra check time, Squadron I, Stage A, and so on for every stage of every squadron to total amount of extra check time for Squadron III.

95 thru 118	UDT 1-A	Used dual time, Squadron I, Stage A, and so on for every stage of every squadron to total amount of used dual time for Squadron IV.
119 thru 149	EC 1-A	Number of extra checks Squadron I, Stage A, and so on for every stage of every squadron to total number of extra checks for Squadron V .
150 thru 154	A 1	Accidents in Squadrons I thru V.
155 thru 159	V 1	Violations in Squadrons I thru V.
160 thru 164	NA 1	Near-accidents in Squadrons I thru V.
165	ET 5-A-3	Extra time in Squadron V, Stage A-3. (Should properly have been in 73, and all others followed from there on as above.)

These items were coded at Tulane University under my immediate direction, to be punched into Hollerith cards, first of all with a view to combining all the items which could be used in predicting ultimate passing or failing before the record was complete. Some of the items contain information that is intrinsically non-quantitative. For example, item 26 (Religion) is the religion which the student mentioned as his own. These answers were classified as jewish orthodox, jewish non-orthodox strict, jewish passive, catholic strict, catholic non-strict but active (this class was empty), catholic passive, etc. Obviously there is no way of arranging these classes in linear order with respect to "degree of religiosity," or with respect to any other characteristic that gives rise to a rectilinear series. For example, a strict protestant might resemble a strict catholic or a strict but non-orthodox jew more closely than he would resemble a passive protestant in so far as his attitude toward religious issues may be concerned. Likewise a passive or nominal catholic might resemble a passive or nominal protestant or a passive jew more closely than he would resemble a strict catholic. However, we did not prejudge this question. Taking all categories separately, we followed the coding procedure of Edgerton and Wherry. First, we took the ratio of passers to total membership in each religion group. From each of these ratios we subtracted the smallest ratio in the series, thus transferring the arbitrary origin to the religion class which gave the smallest ratio. Next, we divided these remainders by the largest ratio, and multiplied the quotients by 9, rounding off to the nearest whole number. This procedure gives the probability of passing, measured from an arbitrary origin, in arbitrary units for each religion category. The numbers which express this result are then correlated with the pass-fail criterion according to the usual Pearsonian formula. The advantage of choosing an arbitrary origin and an arbitrary unit in this manner lies in the fact that the results can be expressed in numbers of one digit and thus simplify coding; and that these numbers are cardinal.

It is unnecessary to our immediate purpose to specify the weights which belong to the various religion classes; they have been made known to the proper authorities.

On the Validity of Wherry's weighting-procedure[5]

Certain friendly critics have questioned whether it may not be illicit to use Wherry's procedure for deriving the weights assigned to non-measurable variables in a multiple regression equation, and to use the same weights for determining the coefficient R of multiple correlation between the combination of weighted test-items and the trait to be predicted, which in this instance is the probability of passing the aviational training course.

I also had a similar misgiving before I had analyzed the problem, but of course resolved it before we adopted the method.

There is only one function for any coefficient of correlation to perform. That function is to aid in description. Its scope is limited by the implications of the defining-equation

$$R^2 = (\sigma^2_y - s^2_y)/\sigma^2_y$$

in which σ^2_y is the variance between the N measured values of Y and their common mean $\bar{Y}$, while s^2_y is their variance from their values calculated from a regression-equation. In this special example the equation has the form

$$\bar{Y}_x = \bar{Y} + (B_1X_1 + B_2X_2 + \ldots\ldots + B_nX_n$$

in which X_1, X_2, ... X_n are the subject's scores in the n different testing traits, and B_1, B_2, ... B_n are the several weighting-factors. These weights are so chosen as to render s^2_y a minimum (for any equation of the same form as this), and thereby to render R^2 a maximum. If the weights are improperly chosen, then s^2_y is not minimized, and R^2 could be enlarged if a better weighting-procedure were employed. The usual method of deriving the proper weighting factors begins with the calculation solely as an intermediate step of the "partial" coefficient of correlation between each testing trait and the trait to be tested. From these n x (n - 1) partial coefficients the weights for each test item may be derived. This procedure like Wherry's, minimizes the variance s^2_y and maximizes R. But since R is used only to express the shrinkage of variance and therefore the shrinkage of standard error between test-prediction and the outcome of training, it makes no difference by what procedure these weights are derived, so long as they minimize the variance. This variance-minimizing combination of weights, however derived, is unique: i.e., every B-factor obtained by one variance-minimizing method has the same value that it would have if it were obtained by any other. Dr. Wherry has presented an algebraic proof that his weighting procedure minimizes the variance. Therefore it is interchangeable with the usual procedure.

[5] Author's note inserted in 1943.

Certain critics questioned, second, whether the weights which belong in this regression equation would probably shrink if the equation were used on another sample of students selected by the same sampling procedure from the same parent-population. This objection is irrelevant to any uses of the information that we proposed to make. Our principal aim was to show what can be done with this type of material. If it were decided to use the questionnaire without alteration -- and it was not -- we would have undertaken to build up a continuing experience-table, such as life-insurance companies build up and such as both our armed forces and the CAA have since been constructing. In so doing, we choose a set of tests which prove to have a usefully high predictive value in one sample; we administer the same tests to a second sample; we compute a new regression-equation for the second sample; we test the two samples for homogeneity in respect to the criterion-trait; we correct if necessary for lack of homogeneity and work out a third regression-equation for the two samples combined; we so proceed until the expanded experience-table has become stable; but we do not cease to experiment as our "experience" is later enlarged.

It usually happens that as the size of the sample, or the samples of samples, increases, the predictive value of the test-battery shrinks, because none of the individual samples perfectly represents the parent-population. However, as Dr. Wherry has pointed out, all these coefficients, and the weights derived from them, have been pre-shrunken by a formula intended to correct for an uncertainty that attends the introduction of every new testing trait into the battery. For when we include another testing trait, we produce two changes. First, we eliminate that part of the total variance that results from our assumption that the B-weight appropriate to the new testing trait is "really" zero. (We presuppose this value if we disregard the test-item.) But while we are eliminating this part of the variance we are otherwise adding to it, for the inclusion of every new test increases the variance due to uncertainties of sampling, since none of the test-item weights is perfectly determined for the whole population in the limited sample that we have. Wherry's procedure was designed to balance the known reduction of variance from the one source against its most probable increase from another source. It requires one to desist from adding more tests when the two effects became equal. If the formula were perfect, then as the sample-population grows, the values of each of these weighting factors and of the correlation-coefficient R would be just as likely to swell as to shrink. It is not urged that the formula is perfect, although it remains to be shown that a better formula has yet been devised. Since, however, it was not proposed to employ the tests in the manner which their critic described, the question of most probable shrinkage becomes hypothetical.

Pre-selective items

Our first and principal problem was to select all those items which could have been used in selection before the student began training at the station. They are indicated by asterisks (*) in the list, from which are excluded those possibly useful items for which the coefficient of correlation with the pass-fail criterion was less than twice its own standard deviation. The test items which failed this standard were as follows:

Age
Previous military experience
Examiner's estimate of heredity
Bernreuter B1N (neurotic tendencies)
Bernreuter B2S (self-sufficiency)
Bernreuter B4D (dominance)
Morgan's test

The items which might conceivably lend themselves to pre-selection, and which also satisfied the standard of admission are shown in Table 12, along with the shrunken multiple coefficient of correlation R which results from their inclusion, one by one, in the test-battery. The rule of inclusion is whether by including the item one increases the predictive value of the test more than one increases the sampling error. The criterion is implied in Wherry's test-selection procedure. These calculations were made under Dr. R. J. Wherry's personal direction, and he is responsible for the result. Table 12 is copied from a memorandum which he sent to the present writer for the Committee in July, 1941.

TABLE 12

	Variables selected	Average N	Shrunken R
26	Religion	254	.2632
30	Major occupations	254	.3153
42	Bernreuter 122	270	.3628
34	Bernreuter 11	291	.3940
33	Fitness, examiner's judgment of	273	.4197
32	Previous training in aeronautics	274	.4372
22	Education	290	.4503
38	Bernreuter 63	298	.4604
23	Major interest or field in college	285	.4704
25	Athletics	278	.4804
39	Bernreuter 104	284	.4892
40	Bernreuter 113	290	.5003
35	Bernreuter 18	296	.5049
29	Environment	288	.5071
36	Bernreuter 44	293	.5109
31	Employment during school year	286	.5104

The table shows that adding the various items to the test battery in the order of their contribution to multiple prediction increases the shrunken coefficient of multiple correlation R until we reach item 31 (Employment during school year). Although this item is positively and significantly correlated with the pass-fail criterion, nevertheless, it is so highly intercorrelated with more highly predictive items that if one makes proper use of them, very little is left for item 31 to measure. Hence, if we had put it into the battery, we should have increased the sampling error by more than we should have increased the accuracy of prediction from the multiple regression equation.

Thus, Dr. Wherry's procedure yielded 15 useful items which, when most advantageously weighted yielded the following multiple regression equation, in which Y denotes the probability of passing the course, and the subscripts to the various X factors denote the nominal number of the test as indicated in Table 12.

The Pearsonian coefficient R of multiple correlation between these 15 test items and the pass-fail criterion is of the order of 0.51. The multiple regression equation reads as follows:

$$\begin{aligned} X_c = -1.5729 &+ .0389X_{22} + .0276X_{23} + .0242X_{25} + .0388X_{26} + .0441X_{30} \\ &+ .0270X_{32} + .0321X_{33} + .0507X_{34} + .0358X_{35} + .0216X_{36} \\ &+ .0623X_{38} + .0246X_{39} + .0361X_{40} + .0203X_{42} + .0093X_{29} \end{aligned}$$

The examiner omitted some of the data on some subjects, so that they had to be restored as well as one could restore them from a fragmentary record blank several years old. The restoration was made by a qualified graduate student in psychology, now in the medical school, who followed strictly the rules prescribed by Dr. Armstrong in every case in which the rules could be applied. A sampling of his restorations was checked by the undersigned, who also was guided by the same set of rules but who disregarded the first inspector's notes. The two sets of scores agreed almost perfectly. That our ratings were based on incomplete information, and might not have agreed with the examiner's ratings if the latter had set them down immediately after the interview is suggested by one or two facts: First of all, the examiner's judgment of the student's "Desire to fly" is correlated 0.45 with eventual success in the earlier work of Dr. Armstrong, Dr. White, and Dr. Kellum. In this exhibit the correlation shrinks from 0.45 to 0.197, say to 0.20, or to less than 45 per cent of its original value. Assuming that the examiners were equally shrewd in appraising this item, one may well surmise that the ratings set forth in Table 12 are based on a greater variety of impressions than were recorded as the basis of the ratings recorded in Table 7. In other words, the written accounts contain only incomplete information; the ratings made by the original examiner were not based solely on the information that he recorded. It is thus evident that if the examiner's judgment is to be considered at all, it is better that he record it immediately after the interview. It loses much of its value if it goes unrecorded until the examiner's memory becomes uncertain, and still more if it is estimated by another from such incomplete notes as these examiners left for posterity.

Although in Table 7 the examiner's judgment of the students' "Desire to fly" has some considerable weight when taken alone, nevertheless when this judgment is estimated by another person from incomplete written records, its value vanishes in predictive importance, when it is considered with other items of biographical information. One reason for this failure lies in the substitution of one mode of estimation for another which we have just discussed. Another reason is that this item is so highly intercorrelated with other test items which gained admission to the test battery before this, and which therefore did so much of the predictional work of the whole battery that little remained for this item to do. But if the original basis of estimation of the student's "Desire to fly" could have been retained, its contribution to total prediction might not have shrunk nearly so much as it did. Corresponding and similar considerations apply to the categories "Inheritance," "Environment," etc. enumerated in Table 7.

It should be remembered, however, that the data on which this study is based were several years old when they were first analyzed; that some of the officers who made the original judgments had been separated from this kind of work for a good while; and that even if they had been available, their problem in restoring the original records would not have been greatly different from our own. This consideration increases the importance of the medical examiner's judgment being recorded without delay. Otherwise, even in reviewing his own work, his procedures are necessarily much like those of an antiquarian.

As we mentioned in the Summary (Part I), some of the students were not rated by the examiner on some of the traits, and the records were so indefinite or incomplete that another person could not construct a substitute rating. We therefore formed two classes of rated subjects. The first class included all those and only those subjects who were rated in respect to every trait that survived Wherry's test selection procedure. Their results are shown in Tables 1-A and 1-B and in Graphs 1-A and 1-B in Part I of this report.

The second class included all the students whether rated in respect to every trait or not. If_any rating was lacking, the gap was filled with the average rating of the whole group for the trait in question. This is equivalent to disregarding the missing trait, for if the scores had been expressed in standard form, the student's rating on every missing trait (in which he was allowed the average gross score) would have been zero, and therefore would have added nothing to his over-all rating derived from the multiple regression equation. These results appear in Tables 2-A and 2-B and Graphs 2-A and 2-B in the Summary (Part I) of this report.

Both tables and both sets of graphs indicate clearly that all the relevant information should have been meticulously collected and recorded for every student; otherwise some data useful for prediction were lost or nullified.

How valuable is this material in selection?

The answer is, "Very valuable if properly used," but the answer demands elucidation. It may be well to begin with a concrete illustration.

Let us restrict our attention to the 208 students for whom every item of information was obtained. Referring to table I-1, p. 2, Part I, of this report, let us suppose that the examiner set the critical score at 0.6, and recommended that all applicants who failed this score be rejected, and that all who attained or exceeded this score be retained for training. Let us suppose also that the training department filed the recommendation and did not examine it until the outcome of training had been completely determined, independently of the tests. The results can be expressed in a 2 x 2 contingency table as follows:

TABLE 13

	A = retained by test score	A = rejected by test score	Total
B - passed the course	$f = 98$	$f = 29$	127
	$f_o = (71)$	$f_o = (56)$	127
	$\Delta = 27$	$\Delta = -27$	0
	$\Delta^2 = 729$	$\Delta^2 = 729$	-
	$\chi^2_o = 10.27$	$\chi^2_o = 13.02$	-
B' = failed the course	$f = 19$	$f = 62$	81
	$f_o = (46)$	$f_o = (35)$	81
	$\Delta = -27$	$\Delta = 27$	0
	$\Delta^2 = 729$	$\Delta^2 = 729$	-
	$\chi^2_o = 15.85$	$\chi^2_o = 20.83$	-
Total	$f = 117$	$f = 91$	208
	$f_o = (117)$	$f_o = (91)$	208
	$\Delta = 0$	$\Delta = 0$	0

$\chi^2 = 59.97$, $P = (10)^{-15}$

The cell AB of the table indicates that there were 98 students who passed the course who would have been retained by the test if the critical score had been set at 0.6. If the test had had no predictive value, the number of retained passers would have been $f_o = 71$, instead of $f = 98$.[6] Thus the test properly classified $\Delta = 98 - 71 = 27$ passers who would have been rejected by this critical test score if the test had been ineffective. Again, the cell A'B' of the table contains $f = 62$ individuals who would have been rejected by the test and who actually failed the course. If the test had been ineffective the corresponding number would have been $f_o = 35$. Thus the test properly classified $\Delta = 62 - 35 = 27$ failers who would have been retained as having been predicted to pass if the test had been worthless. The value of $\chi^2 = \Sigma(\Delta^2/f_o) = 59.97$, which corresponds to a probability $P = (10)^{-15}$ that the discrepancies between the census and the corresponding numbers implied by the independency hypothesis are due to chance.

[6] See pages 10 ff for explanation.

The contents of Table 13 may be indicated in other ways than this. For example, the Pearsonian coefficient of correlation r between retention and passing can be readily computed by formula 13.13 developed in Yule and Kendall's Introduction to the theory of statistics, 11th ed., London, Griffin, 1937. This formula may be written thus:

$$r = \Delta/N\sqrt{p_1q_1}\sqrt{p_2q_2}$$

in which $\Delta = |f - f_o|$ is the absolute difference between the number f of individuals counted in any multiple class (such as the class AB of retained passers) and the corresponding number f_o implied in the independency hypothesis; while N is the number of individuals in the sample; p_1 is the proportion of individuals retained by the test, $q_1 = 1 - p_1$ the proportion of individuals rejected by the test; p_2 is the proportion of passers in the sample and $q_2 = 1 - p_2$ the proportion of failers in the sample.

In this context r has an important meaning, which so far as I have noticed, has not been mentioned elsewhere. Reverting to Table 13 and considering the cells A'B (rejected passers) and AB' (retained failers) we observe that the independency hypothesis requires 56 individuals in the A'B cell and 46 individuals in the AB' cell. Thus a worthless test would classify 56 failers as passers and 46 passers as failers, making a total of 102 individuals improperly classified. Now the most that a perfect test could accomplish is to reclassify these 102 individuals properly, leaving both these cells empty. But the test that we are considering transfers 27 individuals from cell A'B to cell AB, and 27 individuals from cell AB' to cell A'B'. Thus it properly reclassifies 54 of the 102 individuals who would have been wrongly classified by a worthless test, or by chance. But 54/102 = 0.529, which is almost equal to the Pearsonian coefficient of correlation r = 0.537.[7] Thus we can say that this biographical information has about 53 per cent of the predictive value that a perfect test would have.

The superiority of the statistical weighting procedure over the general impression made on an interviewer by the same kind of data is indicated by the coefficient of correlation which in this instance is 0.53 instead of 0.30 as in Table 4. Hence this procedure is nearly twice as

[7] In general the ratio $r' : r = \sqrt{p_1q_2 \; p_2q_1} : \frac{1}{2}(p_1q_2 + p_2q_1)$ -- i.e., as the geometric mean of the two product-terms is to their arithmetic mean. If the difference $p_2 - p_1$ is very small the ratio approaches unity: otherwise the right member of the proportion should be applied as a conversion-factor. A suitable matrix-table is found in: Johnson, H. M. A useful interpretation of Pearsonian r in 2 x 2 contingency tables. Amer. J. Psychol., 1944, 57, 236-242.

effective as the other. Thus an argument which is often used by professional psychiatrists, in favor of allowing them to combine the evidence subjectively, and render the final dicision, is not corroborated in these findings.

So much for the value of biographical information of the kind which we have described, in the segregation of course-passers from course-failers. As we have noted, its value depends partly on the attrition-rate. If this is low, chance or a worthless test will classify most students correctly, so that correspondingly little remains for a useful test to do. Had the value of this inventory been known at the time these students were being considered for admission at Pensacola it doubtless would have been applied. Had it been applied it would have greatly reduced the expense unnecessarily incurred in partially training many inept candidates. Moreover, it would have made more quickly available the services of competent aviators through saving of time wasted in attempts to train incompetents and in awaiting their elimination. Now that the Navy has managed to reduce its attrition-rate by half, the value of this particular inventory for selection has been reduced, most probably by 21 per cent.

It is not to be supposed that this inventory is the best that can be constructed. Its administration requires personal interviews with specially trained officers who combine the skills of psychiatrist, flight surgeon, and father confessor. Their number is limited as well as their time. It is probable that self-administering questionnaires may have to be chiefly relied upon, though supplemented by other information. Two such tests have been constructed-- one by E. L. Kelly, the other by J. W. Dunlap and G. R. Wendt with the enthusiastic cooperation of several members of this committee. This is not a suitable time and place for comparing their merits. Both armed forces and the CAA are already acquainted with them.

But there is biographical information quite apart from personal histories, etc., which promises to be useful, and which can be added to the latter before the students are sent to the naval air station. One of the best potential selectors contained in the records of these classes at Pensacola is the time-investment which the student required for completing certain ground-school courses. But it could not be used because these courses were given at the air station concurrently with flight training. By the time the ground-school record was completed the fate of nearly all the students would have been decided by their performance in flight training. But it is proposed to give some of these critical courses, along with a few hours of pre-primary flight instruction, at the so-called distribution centers or elimination bases. This information should certainly be added to the rest of the students' records and its selective value determined.

A word may be added about a class of tests called (in bad English) "psychomotor." Their selectivity as compared with that of the best available combinations of "paper-and-pencil" tests combined with a good biographical inventory is not great. It may turn out that they can be usefully added to the battery. But it is not easy to get the competent personnel to administer them, and to insure that the mode of administration is kept uniform. It remains to be shown that it is more expedient to employ them than try to improve biographical inventories and certain pencil-and-paper tests with which the members of this committee are acquainted.[8]

Correlation of test items with other criteria

We shall not burden the present report with an account of the many relationships which we examined and found to be too weak to be useful.

One or two other special phases of this investigation, however, are peculiarly interesting. Among the classes which we considered, about 16 per cent of those who failed, or about 6 per cent of those who entered, were washed out before they had completed 15 hours of instruction: i.e., either before they had taken their first solo check, or at this check without benefit of much extra time. On the surface, this looks as if the training department believed that if a student learned slowly in the beginning stages he would probably learn slowly through the later stages or washout. Whether this assumption was verbalized or not, it was not borne out in the records of those who passed.[9] It is well to remember that it was not a generalization from experience, because the experience was curtailed when such men were eliminated: it was an assumption. A similar assumption prevailed at least recently, in another of the services. It may be costly if retained.

However, the time may have arrived when training time must be economized if our planes are to be manned by the time they are needed. Hence,

[8] This assertion is made of the sample population studied and its scope is restricted thereto. Moreover, the author sees no reason for modifying his interpretation. However, it should be noted that since the report was written the Army Air Forces found it practicable to administer certain psychomotor tests, and demonstrated them to have considerable selective value, over and above the contribution made by pencil-and-paper questionnaires. Their sample populations, however, were not homogeneous with this, and it is not easy to make such allowances as are necessary for direct comparison. The whole problem of constructing a test battery is somewhat complicated. For one reason, the order in which component tests are admitted may be important. It should not surprise us if a psychomotor test should be found which was more selective than the best item in our list. Once it is admitted the work that remains for the other component tests to do is correspondingly diminished. Although the question is too complicated to be discussed in detail here, this author sees no incompatibility between the two sets of findings.

[9] This assertion does not bear on another important question: namely, whether students who, instead of being eliminated during the early stages are allowed extra time, are unusually likely to fail in the later stages. In this sample nearly all the failures occurred during primary training. Recent studies made by the RAF, since this report was prepared, indicate that the proportion of student pilots with medium or superior accomplishment in later phases of training, and the proportion reaching operational training, are greater for student pilots who solo early than for student pilots requiring additional hours of dual instruction. (See AAF Aviation Abstract Series, No. 86, 1943.)

it may be important to detect slow learners before they reach the stage of basic training. We therefore asked whether any of the items in this inventory which were available for pre-selection could be used to predict the total extra time of those students who completed the course. The relevant items were 22, 23*, 25*, 26*, 29*, 30, 32, 33*, 34, 35*, 36, 38, 39, 40, 42, as shown in Table 12, p. 32. The six items indicated by asterisks gained admission to the test battery by Wherry's criterion, and be computed a multiple regression based upon them from which to estimate the total extra time. The results were a bit disappointing. Only 130 students' records were complete with respect to the test variables. The coefficient of multiple correlation R = 0.32. The most practicable double dichotomy yields Table 14.

TABLE 14

MULTIPLE TEST SCORE

	A = at least -60	A' = less than -60	Total
B = less than 14 hours total extra time	f = 14	f = 5	19
	f_o = (8)	f_o = (11)	(19)
	Δ = 6	Δ = -6	0
	Δ^2 = 36	Δ^2 = 36	
	x^2c = 4.50	x^2c = 7.20	
B' = not less than 14 hours total extra time	f = 42	f = 69	111
	f_o = (48)	f_o = (63)	111
	Δ = -6	Δ = 6	0
	Δ^2 = 36	Δ^2 = 36	
	x^2c = 0.75	x^2c = 0.57	
Total	f = 56	f = 74	130
	f_o = (56)	f_o = (74)	(130)
	Δ = 0	Δ = 0	0

x^2 = 13.02, P = $(10)^{-3}$, f = 12, f/N = 0.09

Since the total x^2 = 13.02, x = 3.61, r' = 0.20. This finding is reliable but the test enables one properly to reclassify only 12 individuals, about 20.3 per cent of the 59 who would have been improperly classified by chance. This finding is important chiefly because it suggests that much more can be done with such an enlarged and carefully designed biographical inventory as the subcommittee has since drawn up, and which is now undergoing validation.

Conclusions

It is usual to close such a report as this with a set of recommendations, but these have already been made, and some of them have been anticipated. In comparison with those tests of motor coordination, the acquisition of motor skills, and the like which involved elaborate and expensive instrumentation, and a specially trained personnel, and which have not distinguished themselves by their selective efficiency, and in comparison with time-consuming personal interviews, intended to expose latent psychopathologies, this kind of information has proved to be unexpectedly valuable, and also easy to get.[10] It remains to be shown that the other information is necessary although in the meantime we should seek all that we can get. But if it is true that four fifths of the applicants are now excluded by the stringent medical examinations of the armed forces and the CAA, it is hardly to be expected that many would remain eligible for exclusion by an enlarged medical examination.

[10] Editor's Note. Recent studies by the Committee on Selection and Training of Aircraft Pilots and by the Armed Services clearly indicate that measures other than those obtained through the use of biographical information contribute greatly to the prediction of success in learning to fly. Nevertheless, the author's emphasis on the value of biographical data is of particular significance because of the relative neglect of such predictors in selection research in industry.

PART IV. NOTES ON THE ECONOMICAL USE OF PREDICTORS[11]

Certain facts contained in the accompanying tables and graphs can be more easily and usefully interpreted if they are made somewhat more explicit. For suggestions that led to the following procedure, I am indebted to Captain Lybrand Smith, U.S.N. (Ret.) and to Lt. Comdr. John G. Jenkins, H(V)S, USNR.

Given a set of tests the results of which stand in an ordered series, and which stand in multiple correlation with the probability of passing or failing a training-course: one may arbitrarily choose some score K as critical, and consider how many course-passers and how many course-failers attain or exceed it and how many fail it. This reduces the exhibit to a 2 x 2 contingency table, in which the several classes are (a) test-passing course-passers; (b) test-failing course-passers; (c) test-passing course-failers; (d) test-failing course-failers.

Two sets of questions are proposed. The first set includes these: If we set a value for the critical score K; then in order to get 1000 course-passers,

1.1. How many students must we train?

1.2. How many applicants must be available?

1.3. Under the conditions which exist on a given date, have we enough applicants to justify us in selecting a critical score as high as K?

The second group contains these questions:

2.1. If we admit to training a candidate whose test-score falls within a specified range, what is the probability that he will pass the course?

2.2. How does this probability compare (a) with that of the whole sample of students; and (b) with that of the students in the other test-score classes?

The answer to question 1.3 must of course be derived from other information than these or similar tables contain. The answers to all the others can be found in the tables already presented but may be more readily seen in Table 15 and 16, which are extracted from the former, and which are almost self-explanatory. Since they are used solely to illustrate a procedure, I have based them on the total sample, for some members of which the information was not complete. In this exhibit the correlation between test and outcome of training is considerably lower than for those students for whom we obtained complete information, but the number of students is greater.

[11] This part of the report was originally submitted as an appendix in 1942, and has been somewhat condensed.

The questions which we have phrased presuppose a completed experience, in which the proportion of passers to trainees is known. In this sample this proportion P/N = 308/480 = 64.2 per cent, approximately. Tables 15 and 16 are derived from Tables 2-A and 2-B. The results have been slightly falsified in the tails of the distribution by a process of "adjustment" or "smoothing." It seemed to be desirable for two reasons: first, in any part of the distribution the count itself is attended by an uncertainty which is measured by its variance; in either tail of the distribution, where the proportion P/N of passers to trainees is either quite large or else quite small, the variance of the proportion is about equal to the proportion itself. Hence, as we consider one sample after another, we expect this proportion in either tail to fluctuate considerably, so that a smoothing process represents the most probable values better than does an original census comprising a small sample, such as this. Second, in the original data, near the inflection-point of the cumulative frequency curve, the ratio $\Delta y/\Delta x$ is quite large; but since the range of each test-score class is rather broad, this fact makes the test look more sensitive than it really is. The smoothing procedure would not have to be applied if the sample were very large. Whether it is demanded here may be questioned but the answer does not affect the principle that we are trying to expose.

Table 15 which is derived from Table 2-A answers questions 1.1 and 1.2, and provides information that is necessary though not sufficient for answering question 1.3. In respect to question 1.2 it presents in rather startling fashion some facts which I have not seen in formal presentation before.

Consider, for example, the third line. It tells us that if we set the critical score K = 0.9, then, in order to graduate 1000 pilots we need to train (most probably) 1093 students, as appears in Column E. It tells us also that if we had selected for training only those candidates who attained or surpassed this critical score we would have only 8.5 per cent of the total number who would have failed the course, instead of 35.8 per cent for the sample as a whole. But Column D tells us that if we set the critical score K = 0.9, we must have 11,163 applicants to choose from, in order to get 1093 trainees and 1000 passers. In other words, if we have about 11 candidates who have passed the examining board and the medical tests for each candidate that we select then we can cut the washout rate to about one-fourth its present value. But if there are not enough applicants to permit such a high standard of selection as this to be applied, the finding is not practically important.

Consider next the sixth line of the table, corresponding to K = 0.7. If we had selected for training only those candidates whose scores attained or exceeded this value, we should have got a washout rate of about 15.7 per cent, which is still less than half the over-all rate; we should have needed about 2.7 applicants for each passer, and 2.3 applicants per trainee. Perhaps this standard could have been applied. Of course, we now have tests that are probably better than those; their superiority is masked by a reduction of the rate of failure in the Navy-- due, so I suspect, to the adoption of a new policy concerning the allowance of extra time in critical stages.

TABLE 15

Column	A	B	C	D	E	F	G
Line	Critical Test Score	Number of Qualified Trainees	Number of Qualified Passers	Number of Applicants Required for 1000 Passers	Number of Trainees Required for 1000 Passers	Number of Failers	Number of Failers per cent Number of Trainees
1	1.1	6	6	80,000	1,000	0	0.0
2	1.0	18	17	28,235	1,059	59	5.6
3	0.9	47	43	11,163	1,093	93	8.5
4	0.8	117	102	4,705	1,147	147	12.8
5	0.7	210	177	2,712	1,186	186	15.7
6	0.6	290	231	2,078	1,255	255	20.3
7	0.5	359	268	1,791	1,340	340	25.4
8	0.4	408	289	1,661	1,412	412	29.2
9	0.3	439	299	1,606	1,468	468	31.9
10	0.2	460	304	1,579	1,513	513	33.9
11	0.1	472	307	1,564	1,537	537	34.9
12	0.0	479	308	1,558	1,555	555	35.7
13	-0.1	480	308	1,558	1,558	558	35.8

Table 16 contains the information demanded in questions 2.1 and 2.2. Consider the fifth line in the table, which shows the most probable fate of those students whose test-score falls within the range 0.70 to 0.79. If a student belongs in this test-score class, Column D shows that his chances of passing are 84.3 per cent, instead of 64.2 per cent for the average. Column 4 shows that his chances are therefore about 1.24 times the average, and this ratio, though reliable, is not spectacular. But let us drop to the ninth line in the table. It shows that if a student's test-score falls between 0.30 and 0.39, the probability of his passing is only 50 per cent instead of 64.2 per cent for the average, and 79.6 per cent for the men whose test-scores fall within the range 0.70-0.79. In other words the latter have 79.6/50.0 = 1.6 times the chances of passing as have the former. Moreover this ratio is statistically reliable.

I shall not further elaborate what is obvious in these tables. A few points, however, seem to require stress. First, such tests as we have are reliable indicators of what an individual will do as a student of aeronautics. Second, they are quite reliable indicators of the most probable washout rates within the groups that they can be used to classify. Third, in offering them the psychologists are offering nothing magical; nothing that belongs outside the experience and understanding of any medical officer or training officer who can recall a little high school algebra; nothing except an actuarial experience built up from continuing and current practice. As such the tests have a selective value that can be calculated as precisely as life insurance companies calculate the chances of death; their limits of unreliability can be also calculated with accuracy. Hence, if only we keep our experience up to date, the Government may safely depend on the figures.

Column	A	B	C	D	E
				Number of Qualified Passers per cent Number of Qualified Trainees	Probability of passing as multiple average probability
Line	Range of Test Scores	Number of Trainees who Qualify	Number of Passers who Qualify	(= 100 C/B)	(= ND/P)
1	1.1-	5	5	100.0	1.56
2	1.0-1.09	13	12	92.3	1.44
3	0.9-0.99	29	26	89.7	1.40
4	0.8-0.89	70	59	84.3	1.31
5	0.7-0.79	93	74	79.6	1.24
6	0.6-0.69	80	54	67.5	1.05
7	0.5-0.59	69	37	53.6	0.84
8	0.4-0.49	49	21	42.9	0.67
9	0.3-0.39	31	10	32.3	0.50
10	0.2-0.29	21	5	23.8	0.37
11	-0.19	20	4	20.0	0.31

DEFINITIONS: N = number of trainees = 480; P = number of passers = 308; probability of passing = P/N = 308/480 = 0.642 = 64.2 per cent.

The difference between these controlled, and verifiable or falsifiable predictions and those based on the plausible guesses of benevolent but self-constituted experts in World War I, cannot be too strongly emphasized. These data have limitations on their selective value; but it can always be appraised. And the selective tests which this Committee has developed can be counted upon to save the government tens of millions of dollars a year.

APPENDIX

EDITORIAL NOTE

APPENDIX

EDITORIAL NOTE

In considering the foregoing report for publication, questions were raised by the referee concerning (a) the manner of coding the data; (b) the need for cross validation. Following are comments pertinent to these questions.

The data employed by Dr. H. M. Johnson were obtained on 484 subjects at the Pensacola Naval Training Station, coded, and sent to Dr. R. J. Wherry for statistical analysis. The coding procedure is described in detail on page 29. Item 63 on the Bernreuter can be used as an example of how this coding was done. This item could be answered in one of four different ways. The distribution of these answers in comparison with the Pass-Fail criterion proved to be as follows:

Answers	Fail	Pass	% Passing	% P/7.41	Code
Yes	12	5	29.4	3.97	4
?	6	12	66.7	9.00	9
No	139	262	65.3	8.81	8
Omitted	1	0	00.0	0.00	0

Values ranging from 0 to 9 were used in coding, so that they could be punched in a single column on an I.B.M. card. To code the entries in this fashion the largest entry in the third column was divided by 9, giving a value in the case of Bernreuter Item 63 of 7.41 by which all the others were divided. The same procedure was followed in the case of all other items, final figures being rounded off to single digits.

The question has been raised as to whether it was legitimate to use the code scores obtained by this method to obtain the multiple R between the combination of weighted test-items and the probability of passing. It has been suggested that this procedure gives maximum zero order correlations with the criterion used as the dependent variable in the multiple regression equation. It has further been suggested that the effect of maximizing the zero order coefficient is to increase the value of the multiple R beyond its true value. In this connection, the specific proposal was made that one population sample be used to determine the coding; a second to determine the regression weights; and a third to test the shrinkage.

This matter is discussed in some detail on pages 30 ff. of the report. In an elaboration of this discussion, Dr. Wherry has presented the derivations of the formula employed in the application of his method to demonstrate that it actually performs the function of minimizing the variance and that it is interchangeable with the usual weighting procedure. His discussion of derivations is as follows:

[illegible] ... the application [illegible] ... variables [illegible]

Category	[illegible]	Fail	Total
a		[illegible]	fr_{t1}
b		fr_{q2}	fr_{t2}
c		fr_{q3}	fr_{t3}
.		.	.
n		fr_{qn}	fr_{tn}
Σ		N_q	N

(1)

[illegible] formula [illegible] $r =$ [illegible] K (2)

where M_p = mean of the passes on the multi-categoried variable,

where M_q = mean of the failures on the multi-categoried variable,

where σ_t = standard deviation of the multi-categoried variable,

where K = pq/z (if biserial) or $\sqrt{pq}$ (if point biserial).

Let X_i be the assumed cross score weight for category i in the scatter-gram, then

$$M_p = \Sigma X_i fr_{pi}/Np$$
$$M_q = \Sigma X_i fr_{qi}/Nq$$

$$\sigma_t = \sqrt{\Sigma X_i^2 fr_{ti}/N}$$

Substituting in (2)

$$r = \frac{\Sigma X_i fr_{pi}/N_p - \Sigma X_i fr_{qi}/N_q}{\sqrt{\Sigma X_i^2 fr_{ti}/N}} \cdot K \qquad (3)$$

If $x_i = x_i + M$, then

$$r = \frac{\sum (x_i+M)fr_{pi}/NP - \sum (x_i+M)fr_{qi}/NQ}{\sqrt{\sum x_i^2 fr_{ti}/N}} \qquad (4)$$

To maximize r the x_i values [illegible] taking the logarithms of both sides of the equation (4), [illegible]

$$\log r = \log\{\sum (x_i+M)fr_{pi}/NP - \sum (x_i+M)fr_{qi}/NQ\} - \tfrac{1}{2}\log \sum x_i^2 fr_{ti}/N$$
$$+\log N = \log A - \log B + \log N \qquad (5)$$

To maximize r we set the partial derivatives of log r equal to zero, and solve for the values x_i

$$\delta \log r/\delta x_i = \delta \log A/\delta x_i - \tfrac{1}{2}\,\delta \log B/\delta x_i + \delta \log N/\delta x_i$$
$$= \{fr_{pi}/NP - fr_{qi}/NQ\}/A - x_i fr_{ti}/NB + 0 = 0 \qquad (6)$$

Solving all values from 1 to x_k the general solution is

$$x_i = \frac{fr_{pi}/NP - fr_{qi}/NQ}{fr_{ti}} \cdot \frac{BN}{A}$$

$$= \frac{[illegible]}{fr_{ti}} \quad [illegible]$$

$= [illegible]$

$= [illegible]$ pq

$= [illegible]$

$= [illegible]$ (7)

[illegible]

[illegible]

[illegible]

4

[illegible] in the following equation.

$$W_2'(\text{Code}) = \frac{r_{[illegible]}(\text{Element})}{r_{[illegible]}([illegible]) - r_{[illegible]}(\text{Element})} \qquad (9)$$

Multiplying the primed weights by the regression weights to form a composite does not affect them as they are already relative, but merely weights the various tests to produce as little overlap as possible.

Questions concerning cross validation are considered in the body of the report, particularly on page 21. Here evidence is presented to support the point of view that R, as shrunken by the Wherry formula, represents the most probable value of the average R to be obtained from a series of cross validations and thus gives the best prediction of this statistic when cross validation is not possible. Moreover, it is also made clear that the principal aim of the study was to test the applicability of the methods of analysis of biographical material on an available sampling of 434 subjects. Opportunities for cross validation on subsequent samples were not available prior to the completion of this study. It should be pointed out that the Committee on Selection and Training of Aircraft Pilots has recently approved a proposal by the author calling for the application of the techniques of cross validation to biographical items as predictors.