THE HISTORY AND DEVELOPMENT OF THE BIOGRAPHICAL INVENTORY

Prepared

by

National Research Council
Committee on Selection and
Training of Aircraft Pilots

October 1946

CIVIL AERONAUTICS ADMINISTRATION
Division of Research
Report No. 70
Washington, D. C.

National Research Council

Committee on Selection and Training of Aircraft Pilots

Executive Subcommittee

M. S. Viteles, Chairman

C. W. Bray	A. I. Hallowell
D. R. Brimhall	J. L. Holland
P. M. Fitts	E. L. Kelly
J. C. Flanagan	P. J. Rulon

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

October 31, 1946

Dr. Dean R. Brimhall
Assistant to the Administrator
for Research
Civil Aeronautics Administration
Room 3895, Commerce Building
Washington 25, D. C.

Dear Dr. Brimhall:

Attached is a report entitled History and Development of the Biographical Inventory, submitted by the Committee on Selection and Training of Aircraft Pilots with the recommendation that it be included in the series of Technical Reports of the Division of Research, Civil Aeronautics Administration.

The report describes not only an important practical contribution to the selection of student pilots, but represents a significant development in the application of a generally neglected psychological technique in the prediction of work proficiency. There have been significant studies involving the use of biographical data in predicting the success of salesmen. The use of the same method in relation to aviation is an important contribution to psychological methodology as well as to pilot selection.

The series of studies described in the report represents not only a contribution to civil aviation, but helped to lay the ground-work of the cadet selection program of the United States Navy. There is here another instance of the usefulness to the war effort and to other branches of the government of research supported by the Civil Aeronautics Administration.

Cordially yours,

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

MSV:rm

EDITORIAL FOREWORD

The original development of the Biographical Inventory was begun in 1940 as one of the first projects of the Committee on Selection and Training of Aircraft Pilots. Moreover, the development of the Biographical Inventory, together with other work conducted under the auspices of the Committee, in particular that under the direction of Dr. H. M. Johnson,[1] Tulane University, represents one of the first intensive efforts to determine experimentally the value of biographical data in predicting success in flying. In this connection the Biographical Inventory represents an important contribution of the Committee to the selection program of the Naval Air Forces. The version of the Biographical Inventory finally adopted for use in the Naval Aviation Cadet Selection program was based on modifications and revisions of the original inventory, and was developed through the close cooperation of Committee personnel and the staff of the Psychology Branch, Division of Aviation Medicine, Bureau of Medicine and Surgery, United States Navy. The inventory was also made available to the AAF in connection with the development of a selection battery by the Psychology Branch, Office of the Air Surgeon. In addition, it has been employed in a number of research studies sponsored by the Committee, as well as in the extensive CAA National Testing Program, administered by the Committee on Selection and Training of Aircraft Pilots, by means of which applicants were selected for flight training by the CAA War Training Service.

The original development of the Biographical Inventory, and subsequent modifications in cooperation with the Navy, were carried out at Purdue University, under the direction of E. L. Kelly, who, together with D. R. Brimhall, Assistant to the Administrator for Research, Civil Aeronautics Administration, very early became interested in the possibilities of using biographical data to predict flight performance. The present report was prepared by the Editorial Staff of the Committee on Selection and Training of Aircraft Pilots, in particular by E. S. Ewart. Acknowledgment should be made to Commodore J. C. Adams for his cooperation in making available data on the use of the Biographical Inventory in the Naval Aviation Cadet Selection program, and to Lt. H. J. Older, and Lt. (jg) J. B. Carroll for their assistance in the preparation of the section of this report concerning the employment of the inventory in the United States Navy.

---

[1] Johnson, H. M., et al. On the actual and potential value of biographical information as a means of predicting success in aeronautical training. Washington, D. C.: CAA Airman Development Division, Report No. 32, August 1944.

CONTENTS

CONTENTS (Continued)

SUMMARY

In the development of the Biographical Inventory an attempt was made to provide for the collection of a wide variety of biographical and personality data, in order that the value of an extensive sample of such information in predicting various aspects of flight proficiency could be determined. The first experimental questionnaire consisted of 360 biographical and attitude items, 254 items of the type found in personality inventories, 205 words or phrases descriptive of personal characteristics of the respondent who was to check those applicable to himself, and the 400 items in the Strong Interest Blank for Men (Form M). On the basis of criterion groups composed of "safe" and "unsafe" civilian pilots, and "good" and "poor" student pilots in the Civilian Pilot Training program two sets of item weights were developed for the items in this preliminary questionnaire, "J" weights designed to predict "safe" and "unsafe" civilian licensed pilots, and "S" weights, predictive of success in the Civilian Pilot Training program.

Both sets of weights differentiated the criterion groups on which they were based, although validation of the "J" weights on an independent group of civilian private pilots was not possible. The "J" weights did not predict success in the CPT program. The "S" weights were applied to independent samples and yielded a number of relatively high criterion correlations, particularly when correlated against the "skill" factor on the Purdue Scale for Rating Pilot Competency.

Following a number of revisions, the final version of the questionnaire, designated the "Biographical Inventory" for use in the U. S. Navy, was comprised of 105 biographical and attitude items, and 45 personality items. Results of various studies in which the inventory has been used are discussed in this report, particularly with reference to findings bearing upon the reliability and validity of the instrument, and upon the correlations of the inventory with other tests. Although measures of the reliability of the instrument which have been obtained vary considerably, it is concluded that the reliability of the inventory is sufficiently high to render it serviceable for field use.

In regard to validity of the instrument, in most of the studies conducted under the auspices of the Committee on Selection and Training of Aircraft Pilots, the correlations between the Biographical Inventory and criterion measures of flight performance have not been high, being, in general, slightly lower than the criterion correlations of the Test of Mechanical Comprehension, and on the basis of very limited data, somewhat lower than the criterion correlations of certain psychomotor tests. In these studies correlations of all tests with the criterion undoubtedly have been markedly attenuated by the lack of reliability in criterion measures of flight proficiency.

However, although in much of the above research the Biographical Inventory has not predicted measures of success in flight training to a

marked degree, the utility of the Biographical Inventory in the Navy program indicates that the instrument has merit for use in the selection of applicants for flight training, particularly when used as an element of a test battery in a large scale selection program. In this connection, the low correlations between the inventory and tests of mental ability indicated in nearly all studies are of significance. It is suggested that, in future research, attention be given to determining the effectiveness of the Biographical Inventory in predicting those aspects of flight proficiency associated with use of judgment and observance of safety precautions.

THE HISTORY AND DEVELOPMENT OF THE BIOGRAPHICAL INVENTORY

INTRODUCTION

In evaluating suitability for flight training it has long been the custom of flight surgeons to consider information regarding the personal history, background, and interests of applicants, in addition to the results of the physical examination. However, this type of evaluation was not supported to any marked extent by experimental studies designed to determine on factual grounds the relationship between aspects of personal make-up and ultimate success in either civilian or military flying.

This wide use of the biographical approach suggested the desirability of experimental evaluation of biographical items in predicting success or failure as a pilot. The possible usefulness of such studies was also suggested by evidence concerning relationships between biographical items and job success obtained, for example, in investigations involving the selection of sales[1] and other personnel.[2] Moreover, it seemed desirable that such studies be centered on the use of paper-and-pencil instruments in obtaining personal history information both because of the utility of such instruments in collecting standardized information from a large number of cases under controlled conditions and because of the economy of administration of such instruments.

As a result, in 1940, under the sponsorship of the National Research Council Committee on Selection and Training of Aircraft Pilots, a project was undertaken at Purdue University, under the direction of E. L. Kelly, for the development of a questionnaire to be used in the prediction of pilot proficiency.[3] The point of view in terms of which this investigation was formulated was that marked differences in flying habits, evident among groups of pilots all of whom were sufficiently fit to have passed rigid physical examinations, and sufficiently skillful to have passed rigid flight tests, were related to differences in personality make-up which could be determined

---

[1] Considerable work in the selection of salesmen had been carried on by the Life Insurance Sales Research Bureau. See: Kurtz, A. K. Recent research in the selection of life insurance salesmen. J. appl. Psychol., 1941, 25, 11-17.

[2] The usefulness of biographical and personal data taken from application blanks in predicting success on the job is discussed in: Viteles, M. S. Industrial Psychology, New York: W. W. Norton & Co., Inc., 1932, pp. 179-185.

[3] Concurrently with this investigation, research in the same area was also undertaken by H. M. Johnson at Tulane University, under the auspices of the Committee on Selection and Training of Aircraft Pilots. Reference to the results of this investigation was made in later revisions of the Biographical Inventory. See: Johnson, H. M. On the actual and potential value of biographical information as a means of predicting success in aeronautical training. Washington, D. C.: CAA Airman Development Division, Report No. 32, August 1944.

through use of a biographical and personal history questionnaire.[4] In the development of the questionnaire, subsequently called the "Biographical Inventory," an attempt was made to provide for the collection of a wide variety of biographical and personality data in order that the value of an extensive sample of such information in predicting various aspects of flight proficiency could be determined. The inventory was originally designed to predict pilot judgment and other elements of pilot safety. This point of view was subsequently modified to include prediction of general pilot competency in terms of success in flight training which entailed considerably more emphasis on those aspects of flight proficiency identified with "skill."

DEVELOPMENT OF THE ORIGINAL INVENTORY

In order to determine the range of material covered by other investigators and the various types of items employed, a survey was made of other biographical and personal data instruments previously developed. Among the instruments included in the survey were the Woodworth Personal Data Sheet, the Thurstone Personality Schedule, the Bernreuter Personality Inventory, the Strong Vocational Interest Blank,[5] the Humm-Wadsworth Temperament Scale,[6] the inventories employed by the Life Insurance Sales Research Bureau,[7] and the personal data sheet employed by E. L. Kelly in his study of marital compatibility.[8]

Biographical and Attitude Items. On the basis of this survey, and as a result of extended conferences of the staff of the Purdue project, including discussions with flight instructors, a list of approximately 360 biographical and attitude items was developed. Many of the items developed as a result of conferences with flight instructors represented the "hunches" of the flying fraternity as to personal attributes commonly possessed by good or poor pilots. In general, these items were of the type that could be answered by indicating which of several alternatives applied, although two sections of the inventory included statements, the degree of agreement or disagreement with which was to be indicated in terms of a five-point scale from "strongly agree" to "strongly disagree."

---

[4] Kelly, E. L. The relationship of background and personality factors to pilot competency. Washington, D. C.: National Research Council, Division of Anthropology and Psychology, September, 1940. (Progress Report) The description of early work on the Biographical Inventory at Purdue University has been taken largely from this report.

[5] The preceding tests are described in: Garrett, H. E., and Schneck, M. R. Psychological tests, methods, and results. New York: Harper & Brothers, 1933.

[6] Published by the Humm-Wadsworth Personnel Service, Los Angeles, Calif.

[7] See: Kurtz, A. K. Op. cit.

[8] Kelly, E. L. Marital compatibility as related to personality traits of husbands and wives. J. soc. Psychol., 1941, 13, 193-198.

The items in this inventory were grouped under 6 categories or sections as follows:

1. "Parents and Childhood," including educational history and interests.

2. "Present Status," comprising information on marital status, economic status, general accident history, occupation and job history and attitudes, social activities and interests and hobbies, including athletic interests and athletic proficiency.

3. "Your Health," including information on how the subject felt in regard to matters concerning his health.

4. "Your Attitudes," comprising primarily items dealing with generalized attitudes regarding flying and other aspects of pilot behavior.

5. "Your Training and Present Pilot Status," comprising information on flight experience, certificates or ratings held, history of aircraft accidents, other interests and hobbies pertaining to flying, etc.

6. "Your Views About Flying and How It Should be Regulated," comprising items dealing with attitudes toward specific issues related to flying (in distinction to the generalized attitudes with which Section 4, above, was concerned).

These items were presented in a folder entitled "A Research on Pilot Selection - Background, Interests, Attitudes." Also included in this folder were 205 words, or two- or three-word phrases, which were descriptive of personal characteristics, such as "Accurate," "Admits mistakes," "Impulsive," etc. The subjects were directed to "go through the list and put a check to the left of those words which you believe best describe yourself, that is, how you think and act in most situations."

"Personality" Items. In addition to the biographical and attitude items, 254 items of the type found in "personality" inventories were developed. In developing these items it was endeavored to cover the areas sampled by a number of other similar questionnaires as well as to include items considered pertinent to the prediction of success in flying. It was toward the development of items of this type that discussions with flight instructors and other flying personnel were primarily devoted. In addition, the endeavor was made, whenever possible, to phrase the items indirectly, e.g., instead of asking "Do you have spells when you simply can't work effectively?", the item was phrased "Do most people have spells when they simply can't work effectively?". These 254 items were presented in a separate folder entitled "An Analysis of the Normal Personality."

In addition to the biographical items and "An Analysis of the Normal Personality," the Strong Vocational Interest Blank for Men (Form M) also was used in this study. The Strong Inventory contained 400 items.

DEVELOPMENT OF ORIGINAL SCORING KEYS

Selection of Criterion Groups: "Good" and "Poor" Licensed Pilots. As noted previously, the original purpose of this investigation at Purdue University was to determine, through replies made to biographical and personal data questions, certain factors which differentiated "good" and "poor" pilots all of whom had learned to fly skillfully enough to pass the rather rigid flight tests administered by the Civil Aeronautics Authority. More specifically, the purpose was to determine those items to which groups of "good" and "poor" pilots gave differential responses. One approach to the problem of setting up criterion groups of "good" and "poor" pilots (or more properly "safe" and "accident prone" pilots) might have been through the examination of accident records. This procedure, however, did not seem advisable in as much as it would have proved extremely difficult to take into account the influence of extraneous factors such as number of hours flown, types of flights made, planes used, terrain over which the flights were conducted, prevalent weather conditions, etc. Also important was the generally recognized fact that many poor or unsafe pilots have not yet had their first accident. Therefore, it seemed preferable to identify good and poor licensed pilots on the basis of judgments concerning their flying habits made by responsible airport operators.

To this end each of the 500 operators under contract to the CAA in connection with the Civilian Pilot Training program were contacted through the CPT directors at the various schools, and asked to submit the names of two pairs of male licensed pilots. It was requested that the members of each pair be roughly comparable in terms of age, type of certificate held, and number of hours logged, but differing in their tendency to take unnecessary flying risks to the extent that the operator would be willing to rent a plane to one member of the pair but not to the other. The operators were cautioned to be influenced by this consideration only and to disregard other characteristics of the individuals, such as their relative ability to pay or their honesty in keeping accurate account of the time flown. Assurance was given that all information would be kept completely confidential.

Considerable difficulty was experienced in securing this cooperation on the part of the contracting operators, undoubtedly due in part to the reluctance of many operators to brand a fellow pilot as "unsafe." Eventually, however, the names and addresses of 300 pairs of pilots were secured in this manner. Each of the 600 pilots was mailed a set of forms with a personal letter and a printed description of the nature of the research. (In order to avoid the problem of informing certain men that they had been selected as "poor" pilots, the purpose of the investigation was described as the determination of characteristic differences between pilots and non-pilots.) Although the pilots were not required to identify themselves by signing their names to the forms, the tests sent to "good" and "poor" pilots, respectively, were keyed in an inobtrusive manner, although the identity of individual pilots remained unknown. An addressed return postcard was included with the tests by means of which the pilot could indicate that

he had completed filling out the forms. Upon the return of this postcard, one dollar was sent to each pilot who had cooperated.

Two hundred and eight of the 300 "good" pilots and 177 of the 300 "poor" pilots filled out the forms and returned them to the Purdue project. Although a somewhat greater proportion of "good" pilots than "poor" pilots cooperated, the proportion of both groups replying was sufficiently large to represent a usable sample. The 208 "good" or "safe" licensed pilots to whom these responsible operators said they would unhesitatingly rent one of their planes will hereafter be referred to as Criterion Group X. The 177 "poor" or "unsafe" licensed pilots to whom the same operators would not rent one of their own planes will be designated as Criterion Group Y.

Selection of Criterion Groups: "Good" and "Poor" Student Pilots. As the research program developed it seemed desirable also to investigate the personality and background differences among a group of student pilots in the CPT program who experienced little difficulty in learning to fly when compared with a group comprised of student pilots who were extremely slow learners, or who had "washed out" during the training program. In order to implement this extension of the research the CPT directors in 130 institutions were sent a supply of test forms. These CPT directors were asked to cooperate by complying with the following requests:

1. Each director was asked to contact the flight instructors at the local CPT operation, and to obtain from each instructor the name of the best and the poorest of his 10 flight students. (These instructors were also asked to rate the students so indicated on the Purdue Scale for Rating Pilot Competency.)[9]

2. Each director was further requested to call the selected students into his office in order that the research forms could be filled out by the best and poorest students of each instructor. The forms were to be returned to the Purdue Research Foundation.

Only about one-third of the 130 institutions whose cooperation was requested complied, 90 of the 400 pairs of forms sent out being returned in usable condition. The 90 student pilots, considered by their respective instructors as the best of the 10 taught by them, were designated as Criterion Group A. The 90 student pilots, considered by their respective instructors as the poorest of their 10 students, were designated Criterion Group B.

Construction of Scoring Keys: "J" Weights. On the basis of the replies to items in the questionnaires by members of Criterion Groups X and Y ("safe" and "unsafe" licensed pilots), weights were assigned to items in the questionnaires. Weights ranging from -4 to +4 were assigned, the magnitude of the weight for a given item depending upon the difference between

---

[9]Kelly, E. J. The development of "A Scale for Rating Pilot Competency." Washington, D. C.: CAA Division of Research, Report No. 18, July 1943.

the proportions of the criterion groups responding to an item in a specified manner. The sign of the weight was considered positive if the response in question was more typical of the "safe" group, and negative if the response was more typical of the "unsafe" group. The method of determining the item weights was that used by E. K. Strong, Jr. in assigning weights to the Strong Vocational Interest Blank.[10]

Approximately 48 per cent of the 1000 items in the three questionnaires were weighted on the basis of this analysis in terms of the X and Y criterion groups. In as much as these criterion groups comprised, respectively, "safe" and "unsafe" licensed pilots who presumably differed primarily in the amount of judgment exercised and in terms of related factors bearing on the manner in which they piloted a plane, the scoring weights derived from the analysis of the X and Y criterion groups were tentatively designated "J" weights, as possibly predictive of the judgment which a pilot shows in his flying behavior.

Construction of Scoring Keys: "S" Weights. In addition to the "J" weights, item weights for the various items in the questionnaires were also determined on the basis of responses of members of Criterion Groups A and B ("best" and "poorest" student pilots of 90 flight instructors), following the same statistical procedures as were employed in the determination of the "J" weights. Approximately 62 per cent of the 1000 items showed differential responses between the A and B criterion groups of sufficient magnitude to be assigned scoring weights. These weights were tentatively designated as "S" weights, as possibly predictive of skill in learning to fly. Although factors of judgment and emotional stability may be considered by flight instructors in evaluating a student pilot's progress in the CPT course, this designation was made on the basis of evidence from an analysis of the Purdue Scale for Rating Pilot Competency (a criterion instrument which comprised a 14-item rating scale for use by flight instructors) which indicated that elements of performance primarily associated with skill were the major considerations of flight instructors in making such evaluation.

VALIDATION OF ORIGINAL INSTRUMENT

Independent Validation Group. In order to determine the validity of the scoring weights it was necessary to have the forms filled out by a group of pilots other than those making up the criterion groups, and to secure for this new validation group certain external criteria of flight proficiency. To this end the forms were administered to the entire group of student pilots in the Civilian Pilot Training program at Purdue University, the University of Michigan, and to a random sample of about 30 students from Indiana University and Ohio State University. This group, comprised of 88

---

[10] Strong, E. K. Vocational interests of men and women. Stanford University, California: Stanford University Press, 1945, p. 612.

student pilots, was designated Group C. Criterion data on each of these student pilots were obtained in terms of ratings by their flight instructors on the Purdue Scale for Rating Pilot Competency.

Reliability of the Scoring Keys. At the time of scoring the forms for Group C, both "S" and "J" keys were divided in such a way as to secure two scores for each key, one based on the odd numbered items and one on the even numbered items. By this means an estimate of the reliabilities of the scoring keys could be obtained. For this group the correlations between summed scores for odd and even items were .72 for the "S" scores and .68 for the "J" weights. When corrected by the Spearman-Brown prophecy formula, the estimated reliabilities for the complete forms were .84 and .81 for the "S" and "J" weights, respectively. Although not as high as might be desired, it was considered that the weights yielded an acceptable degree of reliability, particularly in as much as the "C" group represented student pilots sampled near the end of the training program and who were, therefore, more homogeneous than an unselected sample could have been expected to be. In view of this fact it was felt that application of the tests to more heterogeneous groups would result in even higher reliability coefficients.

Effectiveness of Keys in Differentiating Criterion Groups. In as much as the "J" weights were derived in terms of Criterion Groups X and Y it might be expected that the "J" weight scores of members of these two respective groups of "safe" and "unsafe" licensed pilots would be significantly different. That this was the case is indicated by the fact that the mean score for Group X was 44.9; the mean score for Group Y, -4.2. The critical ratio of the difference of these mean scores was 13.4, indicating that the difference was highly significant. The "S" weights likewise differentiated between the Criterion Groups A and B on which they were derived, the mean of the A group being 163.6, and of the B group 41.0. The difference of 122.6 between these means yielded a critical ratio of 18.7, indicating that the difference was highly significant. It is of some interest to note that the overlap between the A and B groups was considerably less than the overlap between the X and Y groups.[11]

Application of "J" Weights to Groups A and B. The crucial test of the validity of an instrument is represented by the application of item weights to alien samples, i.e., to samples other than the one in terms of which the weights were derived. As a preliminary step the "J" weights were used in scoring the test instruments filled out by Groups A and B, and the amount of overlapping between these groups which were widely separated in terms of their instructor's judgments of flight proficiency, was determined, as well as the biserial coefficient of correlation between the scores and the tails

---

[11] Although these figures do not represent validity coefficients and should not be interpreted as such, it may be of some interest to note that the biserial correlation coefficient between the "J" scores and the X and Y distributions was .71, and that the biserial coefficient corrected for widespread categories between the "S" scores and the A and B distribution was .84.

of the criterion distribution.[12] It was found that there was a great deal of overlapping between the distributions, the difference between means being only 13.8, which yielded a critical ratio of 2.2 which was below acceptable levels of significance. The corrected biserial coefficient of correlation was .23.

While the "J" weights failed to differentiate or to predict the A and B groups, the ratings of these student pilots by their instructors on the Purdue Scale for Rating Pilot Competency indicated that Groups A and B were not markedly differentiated in terms of items in that scale referring to the judgment and emotional factors presumably used by the operators in selection of Groups X and Y, on the basis of which the "J" weights were derived. In view of this fact, and in view of the fact that A-B, and X-Y groups could be considered incomparable on a priori grounds, this lack of differentiation between Groups A and B in terms of the "J" weights is not surprising.

Application of "S" and "J" Weights to Group C. The crucial test of the validity of the "S" weights and a further test of the validity of the "J" weights were afforded by application of these scoring keys to the test instruments completed by Group C, which comprised the unselected sample of 88 CPT student pilots. The criterion measures available on these 88 student pilots were represented by ratings made by the flight instructors of the members of the group on the Purdue Scale for Rating Pilot Competency.

Item 14 of this scale represents an over-all rating, phrased as follows: "In your opinion, considering skill, emotional stability, judgment, etc., how good an 'all-around pilot' is he likely to become?". The correlations of the "S" and the "J" weights, respectively, with ratings on Item 14 were .45 and -.09. While the correlation between the "J" weights and this criterion is for all practical purposes negligible, it is of significance that the correlation between the "S" weights and the over-all rating is relatively high. In this connection the correlations between the "S" and "J" weights and each of the 14 items on the Purdue Scale are of interest. These correlations are presented in Table 1. Inspection of this table indicates that the "S" scores show correlation coefficients of .40 or greater with Item 1, "Skill"; Item 5, "Feel of Ship"; Item 7, "Tenseness";[13] and Item 14, "Over-all Rating"; and relatively low but positive correlations with most of the remaining items. The "J" weights, however, show low correlations with all items on the Purdue Scale. Of some interest, however, is the fact that the highest correlations (.26 and .23) are with Items 9 and 12 which pertain to "emotional stability," and "judgment," respectively which happen to be exactly the characteristics the "J" scale

---

[12] The correction for discontinuity of distribution was employed. See: Peters, Charles C., and Van Voorhis, Walter R. Statistical procedures and their mathematical bases. New York: McGraw-Hill Book Company Inc., 1940, pp. 393-399.

[13] A high rating on Item 7 indicated lack of muscular tension.

TABLE 1

CORRELATIONS BETWEEN "S" AND "J" SCORES AND THE 14 ITEMS OF THE PURDUE SCALE FOR RATING PILOT COMPETENCY (88 CPT Students)

		"S" Scores	"J" Scores
1.	Skill	.40	.17
2.	Control Handling	.29	.21
3.	Plane Check-up	.25	.18
4.	Ability to "catch on"	.30	.21
5.	Feel of Ship	.47	.19
6.	Respect for Ship	.27	.18
7.	Tenseness	.40	.17
8.	Show off	-.10	.10
9.	Emotional Stability	.23	.26
10.	Confidence	.09	.00
11.	Trying out New Things	.19	.01
12.	Judgment	.11	.23
13.	Self Satisfaction	.25	.19
14.	Over-all Rating	.45	-.09

purported to measure. However, examination of Table 1 indicates that the correlations with other items less directly related to safety, such as "Skill," "Feel of Ship," and "Ability to 'Catch on,'" were not markedly lower.

In addition to the correlations between "S" and "J" weights and individual items on the Purdue Scale, of particular interest are the correlations between these weighting procedures and the composite ratings of three aspects of pilot competency as measured by the scale, namely, "Skill," "Judgment," and "Emotional Stability."[14] The score in terms of Factor I, "Skill," is obtained by adding ratings on Items 1, 4, 5, and 14. Factor II, "Judgment," is obtained by adding ratings on Items 3, 6, 8, 12, and 13. Factor III, "Emotional Stability," is obtained by adding ratings on Items 2, 7, 9, and 10. The correlations between the "S" and "J" weights and each of these factors are presented in Table 2.

Examination of this table indicates that the "S" scores are positively correlated with all three "Factors" or composite ratings, but show the highest relationship with the combination of ratings designated as measurements of the "Skill" component. Factor III, "Emotional Stability," correlates next highest, and Factor II, "Judgment," correlates lowest. The "J" scores show positive but low correlations with all three of the factor scores. Granting the reliability of the "J" weights, a possible explanation would

---

[14] These three aspects of pilot competency, as measured by the ratings on the Purdue Scale, were isolated by a factor analysis of that criterion instrument. Kelly, E. L. Op. cit. (See footnote 9 of this report.)

CORRELATIONS OF "S" AND "J" SCORES WITH FACTORS I, II, AND III

	Factor I "Skill"	Factor II "Judgment"	Factor III "Emotional Stability"
S Scores	.49	.21	.33
J Scores	.23	.24	.19

seem to be that the "J" weights predict some element of flight proficiency not rated (or at best not reliably rated) by the instructors on the Purdue Scale. It seems possible that the rigidly controlled CPT program allowed little opportunity for the student pilots to display, or the instructors to judge, the types of behavior which differentiated the X and Y criterion groups on which the "J" weights were based.

A more informative picture of the predictive value of the "S" weights than is given by the correlation coefficient is afforded by an examination of the scatter diagram in which the relationship between the "S" scores and the ratings of the subjects in terms of Factor I, "Skill," is presented. This scatter diagram is presented in Figure 1. Examination of Figure 1 indicates that none of the nine members of Group C who scored less than 40 in terms of the "S" weights was given a rating by his flight instructor in terms of the "Skill" component above 45, the average for the group as a whole. Furthermore, 8 of these 9 men received "Skill" ratings of 35 or less, i.e., fell one standard deviation or more below the mean of the C group in terms of their "Skill" ratings.

	PURDUE RATING SCALE						
	Low		Skill Ratings				High
Total "S" Score	10-19	20-29	30-39	40-49	50-59	60-69	70-79
180-199			1	1	1	1	1
160-179				6	1	3	
140-159			4	5	2	2	1
120-139			2	2	1		
100-119			2	5	6		
80-99	1	1	5	3	3		1
60-79		1	4	3			
40-59		1	2	3	3	1	
20-39	1	2	2				
0-19			2	1			
-20- -1		1					

FIGURE 1

SCATTER DIAGRAM SHOWING RELATIONSHIP BETWEEN TOTAL "S" SCORES AND RATINGS ON FACTOR I, "SKILL" (88 CPT Students)

It should also be mentioned that further analysis indicated that the "S" scores for the Personal Data Sheet only, correlated .48 with the flight instructor's "Skill" ratings, a value only .01 lower than the coefficient for the total score from all three forms. Five out of the six students with a score below 45 on this form were found to rate in the lowest sixth of the distribution of skill ratings. Although the independent sample on the basis of which the validation was made was not large, the results seemed definitely promising. It was observed that the "S" scores correlated with instructor's ratings on "Skill" approximately as high as scholastic aptitude tests correlate with academic success.[15]

Application of the "S" Weights to Other CPT Samples. During the summer of 1940, two further tests of the instruments were made. The forms were administered to the 50 subjects in elementary CPT training at Purdue University and criterion measurements obtained on the Purdue Scale from each subject's flight instructor. The correlation between the "S" weight, scores and the instructor's ratings was .58. In addition, during the summer session the forms were administered to 50 student pilots in the elementary CPT course at the University of Illinois. The flight proficiency of all 50 of these subjects was rated by a single flight instructor who had an opportunity to observe or to fly with all of the student pilots in the elementary course. The correlation between the "S" weight scores and the ratings of this instructor was .41. It should be noted, however, that in the above situations the tests were administered to the subjects after the initiation of flight training.

In the fall of 1940 the test battery was administered to all applicants for the Civilian Pilot Training program at Purdue University. The conditions of the administration of the battery were such that it appeared to the applicants that whether or not they were selected for flight training would depend, in large measure, on their scores on the tests. Although the number of applicants was considerably larger than the flight program could accommodate in actual fact the test scores were not used at all in selecting men for training. It was possible, therefore, to obtain test results under conditions in which the tests would be used in actual practice, and at the same time to obtain a sample relatively unselected in terms of the tests for use in further investigation of the validity of the battery.

At the completion of flight training the "S" scores of the 50 student pilots were correlated with the ratings given by their flight instructor The highest correlation involving any criterion measure from the Purdue Scale was .11, representing the correlation between the test score and the ratings on the "Skill" component of the scale.

In searching for a reason for this marked drop in prediction efficiency it seemed possible that at least part of the explanation lay in the fact that the groups on which the weights had been derived were administered the tests after flight training had been initiated, whereas the present group took the tests apparently as a condition of admission to the program.

---

[15]Kelly, E. L. Op. cit. (See footnote 4 of this report.)

It might be expected that the answers of the original criterion groups, given under conditions of relatively little pressure would not be similar to the answers of a group which was comparable except for the fact that members felt under pressure to give, not necessarily honest answers to the questions, but answers which seemed to be "right."

As a test of this hypothesis all of the student pilots in this program were contacted and requested to report for a readministration of the tests. A payment of one dollar was offered for this cooperation. Forty of the fifty men complied. Before the readministration of the tests, the situation was explained to these men. They were told that there seemed to have been some conscious effort to "beat" the tests, but that this was perfectly normal and carried no opprobrium. They were requested, however, to give perfectly "honest" answers in this readministration and were assured that their test scores would in no way affect their eligibility for further flight training.

For this re-test, the coefficient of correlation between "S" weight test scores and instructors ratings on the "Skill" component rose markedly to approximately .40, a figure not far below the previous validity coefficients on independent samples. It is recognized, of course, that this increase might be accounted for by the elimination from the re-test group of 10 men from the original sample to whom the test was first administered, or might be due to some other extraneous element, such as the fact that the subjects were paid one dollar for the time spent during the readministration of the forms. Although definitive conclusions cannot be drawn, it would seem somewhat more probable that the increase was due to the major change in administrative conditions, namely the fact that the conditions of the second administration more nearly approximated those under which the test was validated than did the conditions under which the fall group were originally tested. This interpretation lends further substantiation to the accepted fact that the conditions under which a test is validated should approximate, as closely as possible, the conditions under which it will actually be used in practice.

REVISIONS IN ORIGINAL TEST BATTERY

The Bureau of Medicine and Surgery of the United States Navy, through representation on the Executive Subcommittee of the Committee on Selection and Training of Aircraft Pilots was informed of the development of the test battery and the results of the preliminary validation studies. Immediate interest was expressed by the Bureau, and a request made for certain revisions of the battery in preparation for turning it over to the Navy for further validation on samples of Naval aviation cadets, and possible incorporation in the Naval air cadet selection program. In compliance with this request, the following revisions in the instrument were made:

1. The number of biographical and "attitude" items originally included in a folder titled "Background, Interests, Attitudes" was reduced from 300 to 105. This reduction was accomplished by combining certain items and eliminating others which showed extremely low potentialities in the original validation study, or which were of such a nature as to be inapplicable to candidates for training as Naval

flight cadets. For example, items referring to previous job history were eliminated since most applicants were too young to warrant collection of detailed information on job history. Similarly, certain items regarding health which would be covered in the medical examination were eliminated as inappropriate as were items regarding marital status, since at the time the revisions were made cadets were not allowed to marry.

2. The number of items originally included in the "Analysis of the Normal Personality" was reduced from 254 to 150. In effecting this reduction only those items were retained which carried both "S" and "J" weights, and which, on the basis of conferences with personnel from the Bureau of Medicine and Surgery, U. S. Navy, were considered, on authoritative grounds, to show the most promise. This exercise of authoritative judgment was necessitated because of a decision that Part B of the inventory should contain no more than 150 items.

3. The Strong Vocational Interest Blank was eliminated from the battery in as much as its contribution to the validity coefficient when used in conjunction with the other elements in the battery was negligible.

4. The selected items were incorporated in a single test folder which was given the designation, for use by the Navy, of the "Biographical Inventory."[16] Part A of this inventory, entitled "Biographical Data," consisted of 105 biographical items, and items referring to interests and attitudes related to flying. Part B, entitled "How I Feel and Act," consisted of 150 items from the "Analysis of the Normal Personality." Also included as Part C of this inventory, and entitled "Self Description," were 109 pairs of phrases descriptive of personal attributes. In responding to this section, the subject was directed to check the member of the pair which best described him. It was emphasized that checking one member of a pair did not indicate that the subject did not possess the characteristic represented by the other member of the pair. An indication of the phrase which was the more applicable only was required.[17]

---

[16]The designation given this instrument in subsequent research projects under the auspices of the Committee on Selection and Training of Aircraft Pilots was "An Inventory of Personal Data for Prospective Pilots." However, for convenience it is generally referred to as the Biographical Inventory, or the "B.I.," which designation will be used in this report.

[17]Examples of the type of items in this part of the inventory are as follows: 4 (a) collected, (b) concerned about the future; 26 (a) forceful, (b) painstaking. This section of the inventory was developed by Robert J. Wherry and his associates at the University of North Carolina, see: Wherry, R. J. Personality test to avoid guessing the right answers. September 1940. Progress report in the files of the Committee on Selection and Training of Aircraft Pilots.

A further revision in Part B of the inventory was requested following a preliminary analysis conducted by the Bureau of Medicine and Surgery of the responses to the instrument of a sample of Naval flight cadets. On the basis of this analysis the number of items in Part B of the inventory which predicted the criteria of flight proficiency at the 1% level of statistical significance was determined. These items were retained in the final revision and in addition, a number of other non-discriminating items were also included in an endeavor to make the context in which the discriminating items appeared as comparable as possible to the context before this revision of the form. In this revision the number of items in Part B was reduced from 150 to 45.

In addition, Part C, "Self Description," was eliminated. Although on the basis of preliminary analysis this section showed considerable promise, the fact that respondents were required to choose the more descriptive of two alternatives, neither of which in many cases represented a desirable personal characteristic, caused considerable antagonism towards the instrument among applicants for flight training to whom the inventory was administered.

EMPLOYMENT OF THE B.I. IN THE UNITED STATES NAVY[18]

Original Validation. A number of item analyses of this final revision of the inventory were made by the Navy during 1941 and 1942. The most definitive analysis, however, was carried out in the summer of 1942 on the basis of criterion groups obtained by combining all previous cases used in item analyses of this form, and by including additional cases until a total N of 1814 Naval flight cadets was obtained. The key based upon this analysis was put into use September 1, 1942.[19]

Reliability. Two measures of the reliability of this key were obtained by test-retest. The coefficients obtained were .71, based on a sample of 337 cases, and .69, based on a sample of 1989 cases. It would seem reasonable, therefore, to assume that on Naval samples the reliability of the Navy key for the Biographical Inventory is in the neighborhood of .70.

Criterion Correlations. The criterion correlations of the B.I. on five "standard samples" are presented in Table 3.[20] Standard samples are

---

[18]The material for this section of the report was obtained through the courtesy of Commodore J. C. Adams, Division of Aviation Medicine, Bureau of Medicine and Surgery, U. S. Navy. Particular acknowledgment should be made to Lt. Harry J. Older and Lt. (jg) John B. Carroll of the Aviation Psychology Branch for their cooperation in the preparation of this section.

[19]The Navy did not retain the system of differential item weights employed in the original Purdue Study, but soon adopted a system of unit scoring for all weighted items.

[20]In several earlier studies, criterion coefficients of considerably greater magnitude were obtained, some running as high as .60. The samples on which many of these studies were based were poorly chosen, however, since none included all men entering training in a specified period. In addition, some subjects who also had served as members of the original groups on which the item analyses were based were represented, and selected cases tested early in the program also were included.

composed of subjects all of whom entered flight training at approximately the same time, and from which all subjects have been excluded who were represented in the criterion groups employed in deriving the item weights.

In regard to Table 3, the similarity between the values for the first three samples is noteworthy, particularly in view of the fact that the September-October, 1942, sample was scored on a somewhat different key from that used on the two earlier samples. This new key, put into use on September 1, 1942, was based on the analysis conducted in the summer of 1942, as described previously.

TABLE 3

CORRELATIONS BETWEEN THE BIOGRAPHICAL INVENTORY AND OUTCOME OF TRAINING

Sample	N	Criterion Coefficient (Biserial r)
July-November 1941	2355	.30
March-April 1942	1818	.33
September-October 1942	2072	.35
January-February 1943	7744	.21
September-October 1943	6952	.23

All correlations were computed from five letter-grade classes.

The definite and statistically significant drop in the biserials for the 1943 samples cannot be fully explained, although a number of recognizable factors may have been operative. While the first 1943 groups were presumably selected by essentially the same standards as the late 1942 group, an element of "negative selection" entered into the situation in as much as this group represented part of a backlog from which a majority of the men with high MCT scores had already been drawn. Not only would this factor lower the correlation coefficient by curtailing the range of scores, but it also produced a group which was rather poorly qualified for flight training. Furthermore, the interest and motivation of this group may have been lowered by the knowledge that other men were placed farther ahead in the program while they had a new three-month academic course to take. Furthermore, most of the September-October, 1943, group were selected after the Test of Mechanical Comprehension cutting scores had been raised in November, 1942. Since the B.I. and M.C. are positively correlated, some curtailment of the B.I. distribution probably resulted.

Another factor attenuating the correlations is the fact that the population of applicants is known to have undergone various changes as the war went on, e.g., age and educational requirements were changed, motivation doubtless changed as the war status of the country changed, etc. These changes rendered the samples to which the test was administered different from the samples in terms of which it was validated. It seems reasonable to assume that this would affect the validity of the B.I. to a greater degree than it would the validity of other types of selection tests, i.e., tests of other than the personal history type, which would not be as sub-

ject to the influence of cultural, social, and economic changes. It should be mentioned that since tests of the Biographical Inventory type are particularly susceptible to such influences, they must continually be subjected to revision to keep up with cultural changes, influences of changes in social motivation, revisions of training systems, changes in criteria, etc.

A further indication of the lack of comparability between the September-October, 1943, group and previous samples is represented by the fact that the incidence of failure for this group was markedly higher than for previous samples. For example, in the first three samples 75% passed, and in the early 1943 sample 54% passed. In the September-October, 1943, group only 30% passed. The failure rate was to some degree determined by administrative policy, being dictated by the relationship between the number of pilots needed in a given period and the available pool of manpower, as well as by other considerations. This variation in the failure rate undoubtedly caused variations in the standards in terms of which flight performance was evaluated as "passing" or "failing," and this fact undoubtedly is responsible for some of the fluctuations from sample to sample in the correlations of the B.I., and other selection tests, with the pass-fail criterion.

Although the Biographical Inventory was expected primarily to predict flight failures, an examination of the degree to which it predicted failure for other reasons is of interest. In Table 4 are presented the correlations between the Biographical Inventory and various reasons for failure. It will be noted that the highest correlations are for the flight failure group, these coefficients being slightly above those for all failures. It was considered that the B.I. also did a fair job of predicting those who dropped at their own request. The coefficients for the other failure groups are lower although positive.[21,22]

Relationship of B.I. Score and Number of Hours in Flight Training. A number of studies have indicated, as might be expected, that the B.I. score is related to the number of hours necessary to complete flight training. An early study of more than 6000 passers indicated that those graduating most rapidly averaged half a B.I. letter grade[23] higher than did those passing

---

[21]It will be noted in Table 4 that for the September-October, 1943, sample coefficients on "all other failures" and "all failures" are not given. It was not possible to compute these coefficients since, because of the deselection of cadets in this sample by special directive, it was impossible to identify a clear-cut group embracing "all failures." It will be remembered (see Table 3) that the correlation between the B.I. and the pass-fail criterion was .23. Even this coefficient, however, was influenced by the absence of data on outcome of training for 16% of all entrants.

[22]Unfortunately, the criterion correlations of the MCT and the Navy "Personnel Test" (PT) on Navy samples for comparison with those on the B.I. are not at present available.

[23]Biographical Inventory scores frequently are presented in terms of letter-grade ranges from A to E, A representing a high or desirable grade on the inventory.

TABLE 4

B.I. CRITERION CORRELATIONS BY REASON FOR FAILURE

Each population = passers plus designated failure group

Population	Passers	Flight Failures		Dropped at own Request		Ground School Failures		Killed		All Other Failures		All Failures	
	N	r	N	r	N	r	N	r	N	r	N	r	N
Early study: all men entering training before July, 1941	3027	.40	(589)	.32	(110)	.28	(165)	.13	(38)	.10	(113)	.36	(1015)
Controlled Samples													
July-November 1941	1761	.31	(452)	.24	(56)	.11	(45)			.09	(42)	.30	(595)
March-April 1942	1371	.36	(295)	.21	(56)	.15	(24)			.18	(72)	.33	(447)
September-October 1942	1668	.37	(228)	.15	(76)		(13)			.29	(101)	.35	(405)
January-February 1943	4166	.24	(1493)	.18	(931)	.15	(449)			.11	(543)	.22	(3416)
September-October 1943	2112	.29	(1401)	.22	(562)	.22	(630)						

least rapidly. In Table 5 is presented, for the July-November, 1941, sample, an analysis of the mean number of hours in flight training of subjects falling in the five letter-grade ranges of Biographical Inventory scores. It is evident that of the passers, cadets with high scores tend to graduate several hours earlier than those with low scores. It is noteworthy that it is stated that an average reduction of even four hours per cadet would save the government hundreds of thousands of dollars for every 10,000 cadets graduated from flight training.

The picture presented for the failers is less clear. Although the Personnel Test and the Test of Mechanical Comprehension show a regular positive association between test score and hours before failure, the B.I. does not. Even if the small top and bottom groups are excepted, no relationship is discernable, and the difference in means for the C and D groups is not statistically significant. Aside from chance factors, there is stated to be no apparent explanation for the C group remaining longer in training than did the B and D groups.

TABLE 5

B.I. LETTER GRADES AND LENGTH OF TIME IN TRAINING
(July-November, 1941, Sample)

	Average Number of Hours			
	Passers		Failures*	
B.I. Letter Grade	Mean	N	Mean	N
A	209	117	88	9
B	211	634	66	79
C	214	598	76	94
D	216	360	65	98
E	218	46	80	12
Total Group	213	1755	70	292

*Data are not available on about half of the failures in this sample.

Interrelationships Between B.I., M.C., and P.T. In Table 6 are presented the intercorrelations based on these samples, between the Biographical Inventory, Test of Mechanical Comprehension, and Personnel Test, the three tests in the paper-and-pencil battery. The Ns in all cases were greater than 1000 cases. While none of the coefficients is high, it is noteworthy that the correlations between the P.T. and the B.I. are in the neighborhood of zero, markedly lower than the correlations between the P.T. and the M.C. The correlations between the B.I. and the M.C. are only slightly lower than the correlations between the P.T. and the M.C.

Development of Further Keys for the Biographical Inventory. As the use of the instrument continued, the development of three separate keys for the Biographical Inventory was considered desirable primarily for ad-

TABLE 6

INTERCORRELATIONS AMONG THE B.I., M.C., AND P.T.

	Samples		
	July-November 1941	March-April 1942	Sept.-Oct. 1942
P.T. and B.I.	.04	.05	.01
P.T. and M.C.	.33	.30	.29
B.I. and M.C.	.28	.25	.20

ministrative reasons. It was believed that if several keys were used, prospective applicants would be less likely to profit from the leakage of information about the appropriate answers than if a single key were employed.

In line with this development, members of the Aviation Psychology Branch of the Division of Aviation Medicine, Bureau of Medicine and Surgery, had been speculating concerning the possibility that the validity of B.I. items might vary with the pattern of abilities revealed by the A.C.T.[24] and M.C. It was hypothesized, for example, that success in aviation training of individuals with high ability might be associated with responses to the B.I. different from the responses of individuals of low ability.

Separate item analyses of the B.I. were made on five groups of cadets, representing five levels of ability as defined by the A.C.T. and the M.C., and five separate scoring keys were developed. Each of the scoring keys contained relatively fewer items than did the original key based on a sample undifferentiated in terms of ability. A small number of additional scoring weights were assigned to each key on rational grounds, or on the basis of the combination of probabilities for several alternative responses of an item. For example, if for the item "How far in school did your father go" the response "some college" had a positive weight, a positive weight was also assigned to the extreme response "college graduate" even though the latter did not show a significant difference in the item analysis, but providing the non-significant difference was in the proper direction.

Cross validation of these keys on independent samples, similarly differentiated for ability, indicated that while the criterion correlations were generally somewhat higher on a validation group comparable in terms of ability to the one on which the key originally was derived, none of these validity coefficients was higher than were the criterion correlations for the standard key derived from a sample unselected in terms of ability.

Nevertheless, in order to meet the immediate need for several non-overlapping keys for the B.I., it was decided to extend the keys obtained

---

[24] The Personnel Test was supplanted by a somewhat similar intelligence test, the Aviation Classification Test (A.C.T.), which had been developed specifically for use with Naval aviators.

by the foregoing item analysis procedures through adding a certain number of the item weights from the former standardization of the inventory. In general, only those item weights in the previous standardization which had not reappeared in any of the item analyses of ability groups were added to the keys. After considerable experimentation three keys were developed, Key X designed for high ability groups, Key Y for groups in the middle range, and Key Z for low ability groups. Validation coefficients for these keys on two independent samples differentiated as to ability were, in general, comparable to the criterion correlations obtained through use of the original September, 1942, key. As was the case with the previous study on ability groups, while the criterion correlations were generally higher on a validation group comparable to the one on which the key was derived, the differences in criterion correlations of each key among ability groups were, on the whole, not statistically significant.

Correlations of the B.I. with Combat Criteria. Studies of the value of the Biographical Inventory in predicting success in combat are in progress. Although the results of such studies cannot be presented at this time it is of interest to note that preliminary investigation suggests that the patterns of items and responses which predict success in flight training are not necessarily similar to those which appear to predict combat criteria.

FURTHER STUDIES INVOLVING THE BIOGRAPHICAL INVENTORY

A number of other studies involving the use of the Biographical Inventory have been carried out under the auspices of the Committee on Selection and Training of Aircraft Pilots. These studies employed a form of the Biographical Inventory identical with the September, 1942, Navy revision discussed above, and the Navy scoring weights were, in general, used. In the discussion which follows, data from a number of these studies will be presented. The four projects to which principal reference will be made are:

1. A study conducted at Purdue University in the Spring of 1941, which represented a preliminary investigation of the efficacy of a number of predictors of success in flight training.[25]

2. An investigation of the predictive efficiency of a controlled aviation interview, conducted at four centers during the Spring of 1941. During the course of the study the Biographical Inventory and other selection tests were also employed.[26]

---

[25] Kelly, E. L., and Ewart, E. S. A preliminary study of certain predictors of success in civilian pilot training. Washington, D. C.: CAA Division of Research, Report No. 7, December 1942.

[26] Dunlap, Jack W., and Wantman, Morey J. An investigation of the interview as a technique in selecting aircraft pilots. Washington, D. C.: CAA Airman Development Division, Report No. 33, August 1944.

3. The "Boston-Midwest Study," conducted by the Committee on Selection and Training of Aircraft Pilots during 1941 and 1942, the purpose of which was to provide a field trial for a number of promising techniques for the selection of pilots and to provide further data on certain criterion instruments for the evaluation of flight performance.[27]

4. The CAA-National Testing Service. Between June 20, 1942, and January 31, 1943, the CAA-National Testing Service, set up by the National Research Council Committee on Selection and Training of Aircraft Pilots, supervised the administration, on a nationwide basis, of a battery of tests for screening applicants for primary and secondary flight training in the CAA War Training Service program. The test battery consisted of the Biographical Inventory,[28] a Test of Mental Alertness, a Test of Mechanical Comprehension, and a Test of Aviation Information. (This latter test, however, was used only in screening applicants for secondary flight training.) While the purpose of this project was to screen and certify men for flight training in the War Training Service program, a large body of data for supplementary analysis was collected.[29]

Rather than treating each of these studies as a unit, in the discussion which follows the findings from these investigations bearing upon the reliability of the Biographical Inventory, the relationships of the B.I. to other selection tests, and the validity of the instrument will be dealt with separately.[30]

Reliability of the Biographical Inventory. In the analysis of the data from the National Testing Service an estimate of the reliability of the Biographical Inventory, the Test of Mental Alertness, the Test of Mechanical

---

[27]National Research Council Committee on Selection and Training of Aircraft Pilots. Report on the Boston-Midwest Project. Washington, D. C.: CAA Division of Research, Report No. 52, November 1945.

[28]In this program this test was given the designation "An Inventory of Personal Data for Prospective Pilots," although it was identical with the Biographical Inventory used by the Navy.

[29]National Research Council Committee on Selection and Training of Aircraft Pilots. The C.A.A.-National Testing Service. Washington, D. C.: CAA Division of Research, Report No. 39, November 1944.

[30]Moreover, in the interests of clarity and organization, the presentation of material from the various projects, under each category, will not necessarily follow the chronological order of the studies.

Comprehension, and the Test of Aviation Information[31] was made possible because of the fact that applicants who were not certified were eligible for reexamination after a period of six weeks. Since many applicants took advantage of this opportunity, a sample of repeat cases became available. This sample was divided into two sub-samples in order that an estimate of the stability of the results could be made. These will be referred to as sub-sample A, consisting of 367 cases, and sub-sample B, consisting of 1304 cases. The test-retest correlations for these two groups are presented in Table 7.

TABLE 7

TEST-RETEST CORRELATIONS ON PAPER-AND-PENCIL TESTS USED IN THE NATIONAL TESTING SERVICE PROGRAM

	Sample A (N = 367)	Sample B (N = 1304)
Biographical Inventory	.53	.60
Test of Mental Alertness	.75	.79
Test of Mechanical Comprehension	.70	.74
Test of Aviation Information	.75	.77

If these correlations are to be taken as indices of reliability, the Biographical Inventory can be seen to be the least reliable, since in terms of the data from both samples, the rank order in regard to test-retest correlation is Test of Mental Alertness, Test of Aviation Information, Test of Mechanical Comprehension, and the Biographical Inventory. It should be observed, however, that these correlation coefficients are unsatisfactory estimates of reliability since the only group of applicants for which retest data were available was composed of men who had previously failed and who were sufficiently motivated to retake the examinations in an attempt to be certified. Undoubtedly, some of these men attempted to improve their performance by such measures as revising their responses to B.I. items, changing their work rate on the Test of Mental Alertness, guessing more items on the Test of Mechanical Comprehension and Test of Aviation Information, and through similar devices. Not only, then, are these coefficients undoubtedly attenuated, but their relative values are somewhat suspect in as much as it is uncertain whether the attenuation resulting from the factors mentioned above was comparable for all tests.[32]

---

[31] A brief description of these tests is given in: National Research Council Committee on Selection and Training of Aircraft Pilots. Op. cit. (See footnote 29 of this report.)

[32] It is also of interest to note that for all tests the scores on retest of applicants who had failed the battery and later reapplied were significantly higher than the original scores of this group of original failers. However, in the case of the Test of Mechanical Comprehension and Test of Aviation Information the scores of this group of original failers on retest were not significantly different from the original scores of the total group of applicants.

Another estimate of test-retest reliability of the inventory was obtained in the Boston-Midwest study, although this estimate was rendered somewhat suspect by the relatively long interval (15 weeks) which elapsed between test and retest. These test-retest coefficients were computed in terms of two samples of subjects (student pilots) from the Boston area. The number of subjects in the group first tested in the Fall of 1940 was 86, whereas the N for the group first tested in the Spring of 1941 was 73. The reliabilities of various scores for the B.I. as determined in this investigation are presented in Table 8.

TABLE 8

BIOGRAPHICAL INVENTORY, TEST-RETEST RELIABILITY COEFFICIENTS

	Boston Fall Group (N = 86)	Boston Spring Group (N = 73)
	r_{tt}	r_{tt}
+ *1% Items, Part A	.60	.63
± 1% Items, Part A	.57	.68
+ 1% Items, Part B	.66	.64
± 1% Items, Part B	.64	.62
+ 1% Items, Parts A + B	.54	.66
± 1% Items, Parts A + B	.53	.65

*The designation "1% items" indicates that the individual items summed to yield a score each predicted the criterion groups in the Navy validation study at the 1% level of significance. The plus sign (+) indicates that only positively weighted items were summed, the plus-minus (±) indicates that positively and negatively weighted items were summed algebraically.

It is to be noted that the reliability coefficients presented in Table 8 are not high, although they are, in general, of the same magnitude as those determined on the basis of the National Testing Service data. Further analysis indicated, as was evident in the National Testing Service study, that there were increases in mean score between test and retest. Both of these facts might be explained by the relatively long interval of 15 weeks which elapsed between test and retest. Activities occurring during this interval, particularly flight training, conceivably could so alter a subject's interests and his replies to personal history questions that the reliability of the test, as determined by test-retest, would be attenuated.

In view of this fact, it seemed desirable to obtain another estimate of the reliability of this instrument on the basis of test-retest scores under conditions in which a shorter time elapsed between test and retest.

Such an investigation was undertaken at the University of Rochester, under the direction of M. J. Wantman, in connection with the testing of applicants for the National Testing Service program.

The Biographical Inventory was administered to a group of 23 applicants at the first, and again at the last of a testing period in which a number of tests were administered. (A plausible excuse was given for the readministration of the test so that the motivation on the two administrations was probably generally the same.) The time elapsing between test and retest was four hours.

The reliability coefficient, in terms of test-retest, for this group was .94, the total scores being computed on the basis of "A + 1/2" items. Even though the four hour time interval may have been so short that the applicants recalled some of their previous responses, it may be concluded from this small study that the Biographical Inventory is more reliable than the figures obtained from the National Testing Service and the Boston-Midwest samples would imply.[33]

None of these figures on the reliability of the Biographical Inventory can be considered "best" estimates. It is of interest to note that reliability figures obtained in the National Testing Service and Boston-Midwest studies, which were undoubtedly attenuated, and the somewhat inflated reliability coefficient obtained in the study in which only four hours elapsed between test and retest, effectively bracket the reliability coefficients obtained on Naval samples. It seems probable that these latter coefficients, in the neighborhood of .70, represent the most adequate estimates of the reliability of the Biographical Inventory.

Relationships Between Biographical Inventory and Other Selection Tests. In the analysis of data from the Boston-Midwest Project a rather extensive investigation of the intercorrelations between various tests and variables considered potentially useful in the selection of pilots was undertaken. In Tables 9, 10, 11, and 12 are presented in terms of the Boston and Midwest samples, respectively, the intercorrelations among the Biographical Inventory, the Test of Mechanical Comprehension, the Otis Test of Mental Ability, and also (in Table 12, based on the Midwest summer samples) the Test of Aviation Information. Examination of Tables 9 and 10, based on the Boston data, indicates that in both samples the intercorrelations among the four B.I. scores were high, as might be expected since they are not independent (.77 to .94). The correlations between B.I. scores and scores on the Test of Mechanical Comprehension were moderately high, while lower correlations (.07 to .17) are evident between the former and the Otis Test.

---

[33] In connection with this study it was also found that responses to certain weighted items in the instrument were changed from test to retest with greater frequency than were responses to certain other weighted items. This was particularly true in regard to three items, and if these results were borne out in further investigations on a larger scale it would be suggested that certain items in the test are in need of revision or should be eliminated.

TABLE 9

INTERCORRELATIONS OF PAPER-AND-PENCIL TESTS
BOSTON PROJECT (Fall Group)
(N = 100)

		B.I. ± 1% A	B.I. + 1% A	B.I. ± 1% A+B	B.I. + 1% A+B	M.C.	Otis
1.	B.I.* ± 1% A	--	.93	.89	.88	.33	.14
2.	B.I. + 1% A		--	.81	.91	.35	.16
3.	B.I. ± 1% A+B			--	.94	.29	.16
4.	B.I. + 1% A+B				--	.35	.17
5.	Test of Mechanical Comprehension					--	.49
6.	Otis Test						--

*Biographical Inventory

TABLE 10

INTERCORRELATIONS OF PAPER-AND-PENCIL TESTS
BOSTON PROJECT (Spring Group)
(N = 83)

		B.I. ± 1% A	B.I. + 1% A	B.I. ± 1% A+B	B.I. + 1% A+B	M.C.	Otis
1.	B.I.* ± 1% A	--	.93	.77	.85	.39	-.07
2.	B.I. + 1% A		--	.73	.92	.34	-.07
3.	B.I. ± 1% A+B			--	.80	.43	-.01
4.	B.I. + 1% A+B				--	.43	-.02
5.	Test of Mechanical Comprehension					--	.42
6.	Otis Test						--

*Biographical Inventory

TABLE 11

INTERCORRELATIONS OF PAPER-AND-PENCIL TESTS
MIDWEST PROJECT (Spring Group)

		B.I. ± 1% A	B.I. ± 1% B	B.I. ± 1% A+B	Otis	M.C.
1.	B.I. + 1% A	--	-.02	.90	.11	.34
2.	B.I. + 1% B		--	.41	.11	-.06
3.	B.I. + 1% A+B			--	.15	.28
4.	Otis Test				--	.33
5.	Test of Mechanical Comprehension					--

TABLE 12

INTERCORRELATIONS OF PAPER-AND-PENCIL TESTS
MIDWEST PROJECT (Summer Group)

		B.I. ± 1% A	B.I. ± 1% B	B.I. ± 1% A+B	Otis	M.C.	Aviation Information
1.	B.I. + 1% A	--	.11	.88	-.20	.05	.23
2.	B.I. + 1% B		--	.56	-.05	.23	-.02
3.	B.I. + 1% A+B			--	-.19	.14	.18
4.	Otis Test				--	.01	.30
5.	Test of Mechanical Comprehension					--	.25
6.	Aviation Information						--
	N	43	43	42	42	42	42

From inspection of Tables 11 and 12 it is evident that the scores on Part A and Parts A plus B of the B.I. are highly correlated (.90 and .88) on the Midwest sample. This substantiates the finding from the Boston sample, although it should be noted that the two measures are not independent, the score on Part A contributing to the score on Parts A plus B. (It will be noted that the data on the Biographical Inventory from the Midwest Project include only "plus" scoring procedures,[34] but that the scores on Part B of the inventory alone also are presented.) The correlation between Part A and Part B of the inventory is low on both Spring and Summer groups (.02 and .11).

The correlations between the Biographical Inventory scores and the Otis Test are relatively low, as was indicated in the Boston samples, although the correlations between the B I. and the Test of Mechanical Comprehension are considerably lower than on the Boston samples, particularly when the Midwest Summer group is considered. It will be noted that the correlations between the B.I. and the M.C in the Boston samples are somewhat higher than those obtained in Navy samples, whereas these coefficients in the Midwest samples are somewhat lower. However, the general trend for the correlations between the B.I. and Otis to be lower than the correlations between B.I. and M.C. is in general agreement with the findings on much larger Naval samples where, as will be recalled, the correlation between B.I. and the Personnel Test was extremely low. As will be noted in the following section, the correlations between the Biographical Inventory and the Test of Mechanical Comprehension on the Boston-Midwest samples were not a great deal lower than the validity coefficients of these instruments in terms of these samples.[35]

In Table 13 are presented the correlations between various paper-and-pencil test scores and several physiological and psychomotor tests. These coefficients are based on the Midwest Summer sample. It will be noted that the number of cases on which the comparisons between physiological variables and paper-and-pencil tests were made is extremely small (18-19) although 42-43 cases were available for the comparisons involving the psychomotor tests. Despite the small number of cases, certain of the relationships involving the physiological measures warrant brief mention. It is to be noted, for example, that the correlations of the Biographical Inventory (Parts A plus B) with the Vital Capacity measures are relatively high, and that these physiological variables also are relatively highly correlated with the Otis Test. It is also of some interest that the score on

---

[34] In the "plus" scoring procedure only positively weighted items were considered in arriving at the total score. In the "plus and minus" procedure, positive and negative weights were added algebraicly.

[35] Because of this fact and due to its generally higher criterion correlations the Test of Mechanical Comprehension, rather than the Biographical Inventory, was used in several test batteries, the predictive efficiency of which was investigated in the Boston-Midwest Study.

Part A of the B.I. is positively related to Tidal Air measures (about .22) but is negatively related to score on Part B, the coefficients for the two Tidal Air variables with Part B being -.32. However, none of these coefficients is greater than three times the standard error of a coefficient of zero. Additional data from other samples bearing on these relationships would be of considerable interest.

TABLE 14

CORRELATIONS BETWEEN PAPER AND PENCIL TESTS AND PHYSIOLOGICAL MEASURES
MIDWEST PROOF (Summer Group)

	B.I. Part A	B.I. Part B	B.I. Part A+B	Otis	Mech. Comp.	Aviation Information
Body Surface	.03	-.09	-.01	.04	.21	.01
Vital Capacity	.41	-.42	-.52	.37	.10	.14
Tidal Air	.21	-.32	.06	.20	-.23	.32
Tidal Air/Body Surface	.23	-.32	.06	.19	-.23	.34
Vital Capacity/Body Surface	.42	-.40	-.55	.36	.00	.13
Pulse Rate (lying)	-.03	.12	.03	.56	.01	.53
Two-Hand: trial 6	.15	.33	.26	.18	.17	.23
Two-Hand: mean	.15	.32	.27	.22	.28	.16
Mashburn, time through trial 39	-.22	-.23	-.25	-.26	-.09	-.07
Mashburn, time trials 14-26	-.26	-.16	-.30	-.29	-.06	-.05
Mashburn, time trials 27-39	-.27	-.29	-.33	-.21	.00	-.08
Eye-Hand: Pattern A	-.11	-.13	-.15	-.11	-.55	-.36
Eye-Hand: Pattern B	-.14	-.12	-.18	-.16	-.51	-.09
Eye-Hand: Pattern D	-.06	-.12	-.11	.02	-.55	-.35

Number of Cases (N):

Aviation Information vs. Physiological Tests	18
Aviation Information vs. Psychomotor Tests	42
Other Paper-and-Pencil Tests vs. Physiological Tests	19
Other Paper-and-Pencil Tests vs. Psychomotor Tests	43

The correlations between the paper-and-pencil tests and the psychomotor measures are based on a somewhat larger sample. Scores on the Two-Hand Coordination Test are all positively related to the paper-and-pencil tests, while scores on the Mashburn and Eye-Hand Coordination Tests are negatively correlated with the paper-and-pencil tests. On the latter, psychomotor tests, however, a high score denotes "poor" performance. It is perhaps noteworthy that the Biographical Inventory correlates, in general, higher than the M.C. with the Two-Hand and Mashburn Tests, although the coefficients, ranging in absolute value from .25 to .33, are in no case greater than three times the standard error of a coefficient of zero. The Test of Mechanical Comprehension, on the other hand, does show marked correlation with all three patterns of the Eye-Hand Test. The coefficients,

which range from .51 to .59, are not only statistically significant, but are markedly greater than the correlations between this or any other psychomotor test and the Biographical Inventory.

In connection with the Interview Study and the project at Purdue University discussed earlier, extensive investigation of the intercorrelations of the B.I. with other tests was not made. However, mention should be made of the fact that on the basis of the Purdue sample, the correlations of Parts A, B, and A plus B, respectively, with the median interview ratings were .59, -.04, and .39. In addition to this Purdue sample, in the course of the investigation of the aviation interview, data on prediction instruments and criterion measures were available from samples of student pilots in the CPT courses at Harvard University and at Ohio State University. The Biographical Inventory (score on Part A plus score on Part B) was correlated with individual interview scales, the highest correlations with the B.I. being yielded by the ratings on "Fitness for Flight Training." These coefficients were .24 and .36 for the Ohio State sample (N 26-49), and the Harvard sample (N 50-69), respectively. In terms of both samples the B.I. correlated higher with interview ratings than did the Test of Mechanical Comprehension. It might also be noted that the correlation between the Biographical Inventory and the Personal History Inventory (which was used in connection with the interview technique) was .19 and .35 for the Harvard and Ohio State samples, respectively. The correlation between the B.I. and the M.C. on these two samples was .38 for the Harvard sample and .43 for the Ohio State sample.

Perhaps the best estimates of the correlations between the Biographical Inventory and other paper-and-pencil tests are yielded by the analysis of data from the National Testing Service program because of the large number of cases which were available. In Table 14 are presented the intercorrelations among tests used in the National Testing Service. These correlations are based on the "standard" sample of 55,776 applicants, representing about 90 per cent of the total number of over 62,000 applicants. Subjects were excluded from the standard group if they were applying for other than flight training, if they were retaking the tests, or if their registration information was incomplete.

TABLE 14

INTERCORRELATIONS FOR STANDARD GROUP

	B.I.	M.A.T.	M.C.	A.I.
B.I.	--	.079	.257	.413
M.A.T.		--	.351	.375
M.C.			--	.446
A.I.				--
Mean	8.78	39.22	50.64	86.60
σ	2.55	10.43	7.11	34.03

N = 55,776

*Test of Mental Ability.

It will be noted that the correlation between the Biographical Inventory and the Test of Mechanical Comprehension is not markedly out of line with the coefficients found in the Boston-Midwest study, and is of about the same magnitude as the coefficient obtained on the Navy samples. Also noteworthy is the low correlation between the B.I. and the Mental Alertness Test (M.A.T.), (also found on the Naval samples) and the relatively high correlation between the B.I. and the A.I. Although the coefficients presented above were based on a large number of cases, this group was not homogeneous in respect to previous flight experience which is of particular importance in evaluating the high correlation between B.I. and A.I. in as much as this factor could conceivably affect the scores on both B.I. and A.I. Furthermore, although intercorrelations computed from the samples in each of the four "phases" of the program were similar, small differences appeared which proved statistically significant due to the size of the sample. For example, on the four phases correlations between the B.I. and the M.C. ranged from .23 to .28, while correlations between the B.I. and the A.I. ranged from .37 to .45. These shifts presumably were accounted for by changes in the nature of the sample of applicants from phase to phase, e.g., Phase I contained many more secondary applicants than did the remaining phases.

In Table 15 are presented the intercorrelations among the four tests on the basis of a group consisting of "Passers and Selected Failers." This group can be considered relatively homogeneous with respect to previous flight experience and was also the group for which the criterion measure was the most adequate.[36] It will be noted that the intercorrelations among the tests in terms of this sample are considerably lower for this group than for the standard group presented in Table 10, although the coefficients involving the Biographical Inventory in the two tables are comparable in order of magnitude, the Biographical Inventory correlating highest with the Test of Aviation Information in terms of both samples, and lowest with the Test of Mental Alertness.

Further indication of the variation in intercorrelations as a function of the sample on which the coefficients are based was evident when coefficients were computed separately on groups of subjects of various educational

---

[36] Specifically, this group comprised subjects who failed due to inaptitude, because of failure on the flight test or in ground school or because of the necessity of requiring an excessive number of hours of training before reaching a competent level of performance. It was felt that the indicated discrimination of the tests would be increased if subjects who failed for other reasons (disciplinary reasons, illness or injury, etc.) were excluded. This group also included only passers from schools in which there was at least one failure, in as much as the adequacy of the criterion measures at schools where there were no failers was considered suspect. It was known, for example, that at least one of the flight schools in which there were no failers gave an identical grade, viz., 80, to its entire quota of students. Analysis of data from all applicants also was made, however. See: NRC Committee on Selection and Training of Aircraft Pilots. Op. cit. (See footnote 29 of this report.)

TABLE 15

INTERCORRELATIONS AMONG PAPER-AND-PENCIL TESTS AND INSPECTORS' FLIGHT GRADE

(Group Consisting of Passers and Selected Failers from Flight Schools with Attrition)

		1	2	3	4
1.	B.I.	--	-.080	.132	.257
2.	M.A.T.		--	.254	.290
3.	M.C.			--	.397
4.	A.I.				--
	Mean	8.79	42.04	51.98	84.80
	σ	2.13	8.32	6.17	28.76
	N = 2086				

levels. The total sample was divided into three groups in terms of educational background ("Grade School," "High School," and "College"). It was found that the intercorrelations for the grade school group (comprised by those subjects not having advanced beyond grade school education) were, in general, higher than the correlations for the high school and college groups. The largest difference between test intercorrelations for the grade school group and the total group was in terms of the correlation between the B.I. and the M.A.T., the correlation being .253 for the grade school group and .079 for the total group of applicants. It was suggested that the higher test intercorrelations in the case of the grade school group might be due to the fact that at a low educational level the "ability to take a paper-and-pencil test" is a strong common factor affecting performance on all of the tests.[37]

The correlation of the Biographical Inventory with two other paper-and-pencil tests should be discussed briefly. On the basis of samples of subjects from the Standard Testing Program[38] to whom the tests were administered, the correlations between the Biographical Inventory and the Personal History Inventory (P-H) varied on three samples of approximately 400 subjects each, between .21 and .26.[39] These coefficients are in line

---

[37] National Research Council Committee on Selection and Training of Aircraft Pilots. Op. cit. (See footnote 29 of this report.)

[38] The Standard Testing Program was conducted as a preliminary trial of the tests under consideration for use in the more extensive National Testing Service program.

[39] Kogan, L. S., Wantman, M. J., and Dunlap, J. W. Analysis of the Personal History Inventory. Washington, D. C.: CAA Division of Research, Report No. 42, February 1945.

with the correlation between B.I. and P-H obtained on the Boston and Midwest sample discussed previously. The correlations of the Personal History Inventory with the M.A.T. varied between .01 and .05, with the M.C. between .10 and .19. The correlations between the Biographical Inventory and the Desire to Fly Inventory (D-F) on the Standard Testing Service samples varied between .10 and .22.[40] The correlations between this test (D-F) and the M.A.T. varied between .17 and .21 on the three samples of approximately 450 subjects and the correlations between the D-F and the M.C. varied between .24 and .29. Although none of these correlations is high, the fact that the Personal History Inventory correlated highest with the B.I. is undoubtedly a function of the fact that the P-H and the B.I. both contain biographical and interest items. Although the Desire to Fly Inventory also included biographical and interest items, its low correlation with the B.I., and the fact that it correlated somewhat higher with the M.A.T. than did the B.I., suggests that the items in the D-F may, in general, sample different areas of behavior than do the items in the Biographical Inventory.

In summarizing the material bearing on the relationships between the Biographical Inventory and other tests, the generally low correlation with tests of mental ability (the Otis Test in the Boston-Midwest study, the M.A.T. in the National Testing Service data, and the Personnel Test in the Naval samples), and the somewhat higher correlation with the Test of Mechanical Comprehension evidenced on all samples by the B.I. should be noted. These facts are, perhaps, of considerable significance for test battery construction in view of the relatively higher correlations between the M.C. and tests of mental ability evident in terms of all of these samples. The relatively high correlation between the Biographical Inventory and measures based on the Aviation Interview also has significance in connection with the use of the B.I. in conjunction with the interview, as will be discussed later in this report. The generally positive correlations between the B.I. and the psychomtor tests are of some interest. Although the sample on which they were based was not large, the fact that the correlations of the B.I. with the Mashburn and Two-Hand Tests were generally higher than the correlations of the M.C. with these tests is worthy of mention. In regard to the Biographical Inventory alone, the extremely low correlation between scores on Parts A and B of this instrument is also noteworthy.

Of possibly the most significance, particularly in view of the large samples on which the correlation coefficients were based is the evidence of variation in the magnitude of the coefficients of intercorrelation from sample to sample in the National Testing Service program, even when each of the samples is large. Although these variations can be accounted for by changes in the nature of the samples (e.g., Phase I of the National Testing Service included many more applicants for secondary training, i.e.,

---

[40] Kogan, L. S., Wantman, M. J., and Dunlap, J. W. Analysis of the desire to fly (D-F) inventory. Washington, D. C.: CAA Division of Research, Report No. 50, October 1945.

more pilots having had previous training, than did Phases II, III, or IV) the findings, nevertheless, have important implications. Variations of this type could, for example, account for the marked variation in the predictive efficiency of test batteries from sample to sample which frequently is found. Of further interest is the fact that intercorrelations among the tests were markedly higher for samples of low educational level than for samples of higher educational level. The postulation of the "ability to take a paper-and-pencil test" as a strong common factor affecting performance of low education groups on all such tests also has important implications for test battery construction.

Correlations of the Biographical Inventory with Criteria of Flight Proficiency. The Biographical Inventory was administered to the group of applicants for CPT training at Purdue University in the Spring of 1942. Practically no selection of subjects in terms of this instrument was made, however, in as much as the number of applicants was hardly in excess of the quota. A controlled interview by an interviewing board was also held with all applicants.[41]

The correlations between the scores on the Biographical Inventory and several criteria of flight proficiency were determined.[42] The correlations of scores on Parts A and B of the inventory, and total score, with flight instructors' ratings at solo and at the end of the course; and with ratings by a check pilot at the end of the course are given in Table 16. (Ratings by the flight instructors and by the check pilot were made on Item 14 of the Purdue Rating Scale, which called for an over-all evaluation.)

It will be noted that the results in terms of this small sample indicated that Section A of the inventory was carrying most of the weight of prediction and that, when combined in an additive manner, Section B actually detracted from the predictive efficiency of the instrument. The correlation between Section A of the inventory and check pilots' ratings is of practical, as well as statistical significance, however. As noted previously, in terms of this same sample the correlations between the inventory scores and the median ratings of a panel of interviewers were .59, -.04, and .29 when inventory scores on Part A, Part B, and Total, respectively, were involved. As would be anticipated in view of the relatively high correlation between Part A of the inventory and the Interview ratings, combination of these two measures did not result in a marked improvement of the prediction of the criteria over that afforded by either measure individually. It is noteworthy, however, that when combined with the instructor's rating at solo, which can be considered a measure of proficiency on an actual work sample, the multiple correlation of B.I. score, Part A, and instructors' rating at solo with check pilots'

---

[41] Dunlap, Jack W., and Wantman, Morey J. Op. cit. (See footnote 26 of this report.)

[42] For a detailed description and discussion of this investigation, see: Kelly, E. L., and Dwart, J. E. Op. cit. (See footnote 25 of this report.)

rating was .[illegible] (The [illegible] ratings and instructors' ratings at solo [illegible] .[illegible]) Although the sample on which this investigation was carried out was too small to warrant generalizations, the study yielded some suggestion [illegible] of the [illegible] utility of the instrument.

TABLE 16

CORRELATIONS BETWEEN SCORES ON BIOGRAPHICAL INVENTORY AND CRITERION MEASURES
(Range of N: [illegible])

	Instr. Ratings at Solo	Instr. Ratings at End of Course	Check Pilot's Ratings
B.I. Section A	.29	.[illegible]	.47
B.I. Section B	-.03	.00	.00
B.I. Total Score	.[illegible]	.09	.17

As noted previously, in connection with a study of the Aviation Interview, further data on the Biographical Inventory were available on the basis of samples of student pilots from Harvard University and Ohio State University.[43] Of some interest is a comparison between the predictive efficiency of the Biographical Inventory and the Test of Mechanical Comprehension as determined by the correlation with certain criterion measures employed in the interview study. In Table 17 are given the correlations between these tests and various criterion measures available at Harvard and at Ohio State University and the multiple correlations resulting from the combination of B.I. and M.C. scores.

On the basis of the rather limited data presented in Table 17 it is evident that with the exceptions of the rating scale and time measures on the Harvard sample, the Test of Mechanical Comprehension showed somewhat more prediction efficiency than did the Biographical Inventory. It is also of interest to note that combining the B.I. and the M.C. did not, in general, yield multiple correlation coefficients markedly greater than the higher of the zero order correlations. (As noted previously, the correlations between M.C. and B.I. were .38 and .43 for the Harvard and Ohio State samples, respectively.)

It should be noted in passing that the criterion correlations for various measures based on the interview were of the same general magnitude as the criterion correlations of the B.I. and M.C. Because of the relatively high correlations between the paper-and-pencil tests (particularly the B.I.) and the interview measures it was concluded that "The individual interview is an impractical selection technique since less costly group paper-and-

---

[43] Dunlap, Jack W., and Wantman, Morey J. Op. cit. (See footnote 26 of this report.)

pencil tests predict future performance as well or better and the interview results add little to the total predictive efficiency when combined with such tests."[44]

TABLE 17

PREDICTIVE EFFICIENCY OF THE B.I. AND THE M.C. TESTS IN TERMS OF CRITERIA AVAILABLE ON HARVARD AND OHIO STATE UNIVERSITY

	Harvard (N = 50-69)			Ohio State University (N = 26-49)			
	Pass Fail	Purdue Rating Scale	Time Stage D	Purdue Rating Scale	Time Stage D	OSFI Profile Score*	Camera Criterion**
B.I.	.23	.48	-.35	.01	-.12	-.12	-.12
M.C.	.41	.38	-.30	.32	-.12	.47	.34
B I & M C	.41	.53	-.39	.35	-.14	.53	.41

*The OSFI profile scores represent ratings of flight proficiency made by qualified judges on the basis of examinations of the "profile" of maneuver scores on the Ohio State Flight Inventory. See: National Research Council Committee on Selection and Training of Aircraft Pilots, History and development of the Ohio State Flight Inventory, Part II, Recent versions and current applications. Washington, D. C.: CAA Division of Research, Report No. 51, November 1945.

**The Camera Criterion measurements represented ratings of flight proficiency made by trained observers on the basis of photographic records of flight instruments taken during selected maneuvers which yielded pertinent information on the attitude and performance of the plane. See: Viteles, M. S., and Thompson, A. S. An analysis of photographic records of aircraft pilot performance. Washington, D. C.: CAA Division of Research, Report No. 31, July 1944.

Further evidence on the validity of the Biographical Inventory and other selection instruments was obtained in the Boston-Midwest Project. The correlations between the Biographical Inventory and criterion variables, along with criterion correlations of the Test of Mechanical Comprehension and the Otis Test of Mental Ability on the basis of Midwest and Boston samples are presented in Tables 18 and 19. It is evident that although in general the signs of the coefficients are in the expected direction (positive correlations with pass-fail and negative with the time criteria), none of the variables in Table 18 predicts any of the three criteria consistently over the four samples and that the coefficients are for the most part low. Some of the correlations with the Time measures and Pass-Fail approach the lower limits of statistical and possibly practical significance only in terms of the Boston samples. Although, in general, there are evident no consistent relationships between paper-and-pencil tests and

[44] Ibid.

these criterion measures. It should be noted that the coefficients involving the Test of Mechanical Comprehension are in most cases of greater magnitude than those involving the Biographical Inventory.

TABLE 18

CORRELATIONS BETWEEN PAPER-AND-PENCIL TESTS AND PASS-FAIL, TIME FOR STAGE A, AND TOTAL TIME

		BI+1%A	BI+1%A	BI+1%B	BI±1% A+B	BI+1% A+B	Mech. Comp.	Otis	N
Pass-Fail*	Boston Fall Grp.	.23	.25		.22	.25	-.03	.17	98-105
	Boston Spr. Grp.	.07	.31		.25	.27	.32	.13	82-90
	Midwest Spr. Grp.		.15	.02		.15	.31	.07	139
	Midwest Sum. Grp.		.03	.07		.27	.35	.14	40-47
Time Stage A	Boston Fall Grp.	-.07	-.15		-.14	-.14	-.24	-.17	83-84
	Boston Spr. Grp.	-.28	-.26		-.28	-.27	-.27	-.01	72-80
	Midwest Spr. Grp.		-.09	.00		-.08	-.12	-.04	137-138
	Midwest Sum. Grp		-.06	-.08		-.24	-.35	-.06	40-47
Total Time	Boston Fall Grp.	-.21	-.29		-.18	-.20	-.37	-.10	83-84
	Boston Spr. Grp.	-.32	-.25		-.23	-.21	-.27	-.08	72-80
	Midwest Spr. Grp.		.07	-.14		.01	-.24	-.04	121-122
	Midwest Sum. Grp.		-.07	-.03		-.07	-.17	-.07	35

*Correlations with Pass-Fail are expressed in terms of biserial coefficients. Other correlations are Pearsonian coefficients.

TABLE 19

CORRELATIONS BETWEEN PAPER-AND-PENCIL TESTS AND PURDUE RATING SCALE (Rating on Item 14)

		BI+1%A	BI+1%A	BI+1%B	BI±1% A+B	BI+1% A+B	Mech. Comp.	Otis	N
Stage A	Midwest Spr. Grp.		.15	.25		.22	.24	.31	25
	Midwest Sum. Grp.		.18	-.12		.12	-.04	-.33	9
	Boston Spr. Grp.	.35	.26		.27	.37	.23	-.05	76-82
Stage D	Midwest Spr. Grp.		-.01	.02		.02	.19	.10	39-40
	Midwest Sum. Grp.		.07	.44		.26	.48	-.10	12
	Boston Spr. Grp.	.43	.40		.33	.39	.13	-.16	76-82

Examination of Table 19 indicates that the relationships between the paper-and-pencil tests and the ratings on the Purdue Rating Scale are generally comparable to the relationships between these predictors and the criteria presented in Table 18, although it is perhaps worthy of comment that the Biographical Inventory, Part A, predicted relatively well in the Boston sample, and less well in the Midwest sample, and that the correlations between the paper-and-pencil tests and the Purdue Scale were, in general, higher than were the correlations of these tests with Pass-Fail and Time criteria.

From the analysis of the National Testing Service data, estimates of the predictive value of the Biographical Inventory and other paper-and-pencil tests also were obtained. The cutting scores on the principal battery of three tests (the B.I., the Test of Mental Alertness, and the Test of Mechanical Comprehension) were set at points designed so that failure on one or more of the three tests would occur in the case of 30 per cent of the candidates. The estimate of the percentage to be eliminated, as well as the particular cutting scores for the individual tests, was based on an analysis of previous studies, especially those carried out in connection with the Naval Aviation Cadet Selection program.

Analysis indicated that about 10 per cent of the total group of applicants fell below the cutting score on the Biographical Inventory. However, the Test of Mental Alertness rejected about twice as many applicants as the Biographical Inventory and more than twice as many as did the Test of Mechanical Comprehension, which rejected slightly less than the Biographical Inventory, or about 8 per cent of the total number of applicants. It is of significance, however, that when the data were broken down in terms of educational background of the group of applicants, the Biographical Inventory was found to have eliminated more applicants (about 13 per cent) who had college training than either of the other two tests. With the exception of the Biographical Inventory the proportion of subjects eliminated by each of the tests increased from the college through the high school to the grade school educational level. The B.I. eliminated the greatest proportion (19%) at the grade school level, but least at the high school level (7%).

Two estimates of the criterion correlations of the tests are available from the most adequate sample of subjects from the National Testing Program, i.e., the group of 2086 subjects consisting of passers and selected failers from flight schools with attrition.[45] The correlations between individual tests and the inspectors flight grade[46] in terms of this group

---

[45] For definition of this group see footnote 36.

[46] In case of failure in the flight test the inspector merely indicated the reasons but assigned no grade. In computing these coefficients, therefore, subjects who failed the flight test were given an arbitrary flight grade of zero, in as much as an analysis indicated that this represented the most conservative procedure. See: National Research Council Committee on Selection and Training of Aircraft Pilots. Op. cit. (See footnote 29 of this report.)

were as follows:

Biographical Inventory	.067
Test of Mental Alertness	.067
Test of Mechanical Comprehension	.141
Test of Aviation Information	.108

In considering the correlations between the prediction tests and the flight grade, reference should also be made to the biserial coefficients of correlation between these tests and the Pass-Fail criterion. These coefficients are presented in Table 20 for the same sample under discussion above.

TABLE 20

BISERIAL CORRELATIONS BETWEEN TESTS AND PASS-FAIL CRITERION
(N = 2068)

Passers and Selected Failers with no Previous Flight Hours
from Flight Schools with Attrition

		M_P	M_F	σ_T	N_P	N_F	Bis-r
1.	B.I.	8.8	8.4	2.1	1927	159	.090
2.	M.A.T.	42.1	40.7	8.3	1927	159	.077
3.	M.C.	52.1	49.3	6.2	1927	159	.229
4.	A.I.	85.6	75.4	28.8	1927	159	.174

It will be observed that these criterion correlations are positive, but low. It is of some interest, however, to note that relatively the Biographical Inventory shows less predictive efficiency than either the Test of Aviation Information or the Test of Mechanical Comprehension in terms of both the flight score and the Pass-Fail criteria. A similar situation was evident from the analysis of the total group of passers and all failures (with no previous flight training), the relative predictive efficiency of the Test of Aviation Information, the Test of Mechanical Comprehension, and the Biographical Inventory remaining the same, although in this sample the Mental Alertness Test proved somewhat superior to the B.I. It might be noted, however, that in the case of the criterion correlations for those subjects having previous flight hours, the coefficient for the Biographical Inventory was somewhat higher (.109) than for the group having no previous flight hours, whereas the coefficient for the Test of Mechanical Comprehension dropped to approximately the same magnitude as the Biographical Inventory (.105).

In interpreting these correlations as "validity coefficients" it must be remembered that all of the samples were curtailed at the lower score levels. In the general population which originally took the tests, approxi-

mately 30% failed on one or more of the basic three tests in the battery (B.I., M.A.T., and M.C.), and since these cases were, in general, not certified for flight training, no criterion data on these subjects are available. As a result of this screening the apparent predictive efficiency of the tests as indicated in the present analysis is probably lower than if all cases originally tested had been allowed to continue through flight training. Moreover, since the proportion of failers in the present sample is so small, the product-moment correlations (with the flight grade) were substantially the same for the passers only as for the total population including failers. The criterion correlations in effect represent the relationship between the tests and the flight proficiency of passers falling above the cutting scores in the tests.

A further factor which undoubtedly attenuated the criterion correlations is the unreliability of the criteria of flight performance, namely, the Flight Grade and the "Pass-Fail" criterion, and differences in standards in terms of which flight proficiency was assessed. Although no definitive criterion study was undertaken of the data collected in connection with the National Testing Service, analysis of distributions of flight grades taken from consecutive record cards (for ten flight schools in Texas) indicated that the mean grades for men passing the flight test at respective schools ranged from 72.4 to 92.4. No overlap whatsoever was evident in the distributions for three of the schools. In addition, marked differences in range of grades from individual schools were evident. In one school a single flight grade, 80, was given to all 15 men. In contrast, another school showed a range of grades from 70 to 90. Such variations in means and ranges of flight grades undoubtedly have a limiting influence on the size of correlations between predictors and flight grades, and undoubtedly account, at least in part, for the low criterion correlations found in this study.

In summarizing the analyses bearing on the validity of the Biographical Inventory it is evident that this instrument did not show marked predictive efficiency except in the case of the Purdue sample, which was relatively small. Although the criterion correlations of the Test of Mechanical Comprehension similarly were not predominantly of statistical significance, it is noteworthy that over most of the samples this instrument predicted somewhat more satisfactorily than did the Biographical Inventory. The Biographical Inventory, however, in nearly all samples (including Naval groups) showed lower correlations with tests of intelligence than did the Test of Mechanical Comprehension, suggesting the value of its inclusion in test batteries in which a test of intelligence is incorporated. In this connection it should be noted that the Biographical Inventory has functioned well in the battery employed in the Naval Flight Cadet Selection program in which the criterion correlations have been sufficiently high to warrant its continuance as an integral element in the Navy battery.

Another suggestive finding was the evidence obtained from the National Testing Service data, that a relatively greater proportion of men with college training fell below the cutting score on the Biographical

Inventory than [illegible] on the Test of Mechanical Comprehension and the Test of Mental Alertness. This, of course, does not constitute prima facie evidence that the B.I. is more predictive at the college level than in regard to persons with less advanced educational background. It does, however, suggest the desirability of determining experimentally the efficiency of the Biographical Inventory in predicting ultimate flight proficiency of groups of subjects differing in terms of educational background.[47]

DEVELOPMENT OF A "CIVILIAN KEY" FOR THE BIOGRAPHICAL INVENTORY

As was evident from the discussion in the previous section in all of the situations in which the Biographical Inventory has been employed, subsequent to its original development, the "Navy weights" have been used in scoring the instrument, i.e., the item weights used were those derived on the basis of responses from Naval flight cadets. In view of projected use of the instrument in connection with research on civilian pilots it was thought advisable to make an approach toward the development of a key based on the responses of civilians. As a first step in this direction, item weights were determined on the basis of samples of civilian men who were administered the test in connection with the National Testing Service program, and who later engaged in flight training under the CAA War Training Service program.

This civilian key was constructed on the basis of a sample from the National Testing Service program having no previous flight training or experience, in as much as it was considered that this limitation rendered the group most comparable to applicants for flight training in situations in which the instrument might be used. The criterion of flight proficiency used in the derivation of item weights was dichotomous, the criterion score of any individual being determined merely by whether he passed or failed the flight training course. The sample on the basis of which the item weights were determined consisted of 3151 cases. Of these, 2917 subsequently passed flight training and represented the "passer" group. The group of "failers" consisted of 234 cases who did not pass their subsequent flight training course.

The predictive value of each item in the inventory in terms of the Pass-Fail criterion groups, was evaluated through use of the chi-square test. All such chi-squares were computed in terms of four-fold contingency tables, those items which had more than two possible answers being analyzed by means of as many four-fold tables as there were possible answers. Each contingency table in this case represented a dichotomy on the one axis be-

---

[47] It should be noted in this connection, however, that studies on Navy samples yielded no evidence that the prediction value of the B.I. varied as a function of level of intelligence, as defined by the Personnel or Classification Tests and the M.C., and within the range of intelligence of the sample as defined by these tests.

tween those giving the specific answer in question versus those not giving the answer, and the pass-fail criterion on the other axis.

Those items which gave a p value on the chi-square test of .05 or less were assigned unit weights in the scoring key. If the item was marked by a greater proportion of "passers" than "failers" it was given a positive weight. If the proportion of "failers" marking the item was greater, it was given a negative weight. Those items which yielded chi-squares significant at greater than the level of .05 were not weighted, and contributed nothing to the final score. Analysis indicated that in these terms 52 items in the Biographical Inventory were assigned weights. In as much as only 3 of these items were in Part B of the Inventory, this section was not scored and the 49 weighted items included in Part A of the inventory only were weighted in the final scoring key.

An estimate of the predictive efficiency of this civilian key was obtained by determining the correlation between the score on the inventory and the pass-fail criterion measure on the basis of an independent sample. The sample used represented 974 cases from the Standard Testing Program, and members of which had no previous flight training or experience and including 865 passers and 109 men who failed the flight training course.

The correlation between the total score on the inventory in terms of the civilian key and the pass-fail criterion was .22. It will be noted that this "validity coefficient" is not high, being only slightly higher than the coefficients reported for the National Testing Service samples in which the Navy weights also were employed. It should be remembered, however, that the weights on which this scoring key was based cannot be expected to be as adequate as if they had been derived on an unselected sample. The sample on which this key was derived had already been screened by means of the three tests administered in the National Testing Service, one of which was the Biographical Inventory. Thus, the group on which the key was derived, as well as the independent validating group, represent samples in which the tail of the distribution has been cut off.

The exact amount of this curtailment can only be estimated. Approximately 9 per cent of the applicants for flight training of the group from which the sample used in deriving the weights was taken, had scores below the cutting point on the B.I.[48] Since the scores of men without previous flight training were lower than the scores of the group as a whole, it is reasonable to suppose that the actual curtailment was greater than 9 per cent. Moreover, since there is a positive relationship between the B.I. and the M.A.T. and M.C., respectively,[49] the curtailment which was

---

[48]The sample used in the development of the civilian weights was taken from data from Phase I of the National Testing Service. See: National Research Council Committee on Selection and Training of Aircraft Pilots. Op. cit. (See footnote 29 of this report.)

[49]Ibid.

due to these factors [illegible] reduce the number of individuals who, while exceeding the [illegible] on the Biographical Inventory, had low scores in terms of this criterion. There was a similar curtailment of scores on the validating sample taken from the Standard Testing Service program.

It is of interest to note that only three of the items carrying weights on the civilian key are also weighted on the key, based on a Navy population, used in other studies. Whether or not this is due to the curtailment of the sample or results from the fact that the samples on which the civilian key were developed differ markedly from the Navy sample, cannot be stated. In this connection, however, it should be mentioned that on the basis of the independent sample used in testing the civilian key, the correlation between the pass-fail criterion and the inventory scored with the Navy weights was exactly the same as the correlation between the civilian key and the pass-fail criterion. In both cases the biserial coefficient was .22.

It is evident that the civilian key for the Biographical Inventory as developed on the basis of this analysis described above, cannot be considered adequate for use in the field. The unquestionable lack of reliability of the pass-fail criterion attenuated the criterion correlations of the key. This factor also results in the item weights being less valid than if the criterion had been more adequate.

Nevertheless, in view of the marked limitations of this investigation, the fact that a key could be developed which yielded a criterion correlation on an independent sample that was, although not high, at least statistically significant, is not unpromising. In undertaking further research in this area it will not only be essential that the samples on which the item weights are derived be representative, but also that the criterion measures in terms of which the key is constructed and validated be of sufficient reliability and validity to render possible the development of a key with satisfactory predictive value.

DISCUSSION

In considering the results of the various investigations under the direct auspices of the Committee in which the Biographical Inventory has been used it seems evident that completely adequate estimates have not been made of either the reliability or of the validity of the instrument. The determinations of reliability based on test-retest have been influenced by the length of time between test and retest, excessively long intervals rendering the estimate too low, while the estimate based on retest after four hours undoubtedly yielded an estimate which was somewhat too high. Furthermore, the measure of reliability based on the National Testing Service sample was rendered suspect due to the fact that selective factors determining the nature of sample on which the retest was based and the motivation of subjects in this sample could be expected to attenuate the reliability coefficient. In addition, the estimate of reliability of the original test battery from which the Biographical Inventory

was developed which yielded a coefficient of about .80 was based on a great many more items than were included in the final version now in use, and was computed in terms of the split-test technique which is perhaps not as applicable to an interest inventory of the type represented by the Biographical Inventory as to other types of test with which this procedure is frequently employed.

Under conditions in which recognizably extraneous factors undoubtedly attenuated the reliability coefficient, the correlations between test and retest were in the neighborhood of .60. Under conditions in which the retest followed after a period of four hours, which minimized the influence of extraneous factors attenuating the correlation but which introduced a memory factor which probably rendered the coefficient somewhat too high, the estimate of reliability was .94. In as much as these figures probably represent the limits between which the "true" reliability lies it appears that the reliability figures, in the neighborhood of .70, obtained in analyses conducted by the Navy, represent the best estimate of the reliability of the instrument. Although the reliability of the Biographical Inventory may not be as high as could be desired, nevertheless, it appears sufficiently high to render the inventory practically useful, at least as far as the consideration of reliability alone is concerned.

Information on the critical matter of the validity of the Biographical Inventory in many cases unfortunately is not definitive or complete. In investigations conducted by the Committee on Selection and Training of Aircraft Pilots the correlations between the inventory and various criterion measures of flight proficiency, in general, have not been high. In most of these cases, however, conditions were present which rendered the determination of validity somewhat suspect. In the Boston-Midwest Project and in the studies conducted at Purdue University following the original development of the inventory, the number of cases available was relatively small. In regard to the National Testing Service data, while the number of cases available was extremely large, the correlations were reduced by the restriction of the range of both test scores and criterion measures, which resulted from the pre-selection of the sample by means of a test battery which included the Biographical Inventory. Perhaps the best estimates of the validity of the instrument have resulted from analysis conducted by the Navy, which place the validity coefficients of the instrument perhaps somewhat higher than .30.[50] Even in these studies, however, conditions attenuating the coefficient were recognizable in many instances.

Of major importance (and applying to a considerable degree to the Navy studies as well as to investigations conducted by the Committee) is the fact that the lack of reliability and the general inadequacy of many criterion measures effectively place a practical ceiling above which the criterion correlations of prediction tests cannot rise. The inadequacy of the criterion measures, as noted previously, undoubtedly was responsible for considerable

---

[50] It should be emphasized that the Navy's problem was always that of mass selection. Tests yielding criterion correlations of this magnitude are effective if applied in this connection, although their use in individual counseling would, of course, be highly dubious.

attenuation of the criterion correlations determined on the basis of the National Testing Service data. It is perhaps significant in this connection that in other studies over several samples some of the highest criterion correlations involving the Biographical Inventory were obtained with the Purdue Rating Scale, a criterion instrument which not only has been demonstrated experimentally to differentiate between "good" and "poor" pilots,[51] but which also yields a relatively wide range of criterion scores. On the other hand, however, the Biographical Inventory has not predicted even this criterion measure on other samples although without question the reliability of this criterion instrument varies as a function of the care with which it is administered.

Granting that the inadequacies of criterion measures reduce the criterion correlations of prediction tests, the relative predictive efficiency of the Biographical Inventory as compared with other tests is of significance. It is noteworthy that although, in general, the inventory predicted better than did tests of mental ability, it was somewhat inferior over the majority of samples to the Test of Mechanical Comprehension,[52] and particularly in the National Testing Service analysis, to the Test of Aviation Information. On the basis of the Boston-Midwest data, which admittedly were limited, the inventory predicted somewhat less satisfactorily than did certain psychomotor tests, particularly the Mashburn and the Two-Hand Coordination test, although for none of these tests were the criterion correlations high.

However, in regard to the construction of paper-and-pencil test batteries, it is of importance to note that the correlations between the Biographical Inventory and tests of mental ability were consistently lower than were the correlations between the Test of Mechanical Comprehension and tests of mental ability and that this relationship was also evident in the Navy data. This suggests utility of the Biographical Inventory in test batteries which include a test of mental ability.

Finally, in regard to the general applicability of the Biographical Inventory in the practical situation, it should be emphasized that the inventory has never been employed primarily for the purpose for which it originally was developed and on the basis of the specific frame of reference in terms of which the selection of items originally was oriented. As noted in the introduction to this report, the point of view originally taken was that the inventory should be designed primarily to predict pilot judgment and other elements of pilot safety. It was designed, in its inception, in an endeavor to predict marked differences in flying habits, evident among groups of pilots all of whom were sufficiently fit to have passed rigid physical examinations and sufficiently skillful flyers to have passed rigid flight examinations, in the belief that such differences

---

[51] Kelly, E. L. Op. cit. (See footnote 9 of this report.)

[52] As noted previously, due to unavailability of Navy data, it is not possible to state whether or not this trend was borne out by the experience of the Navy with these instruments.

in flying habits were related to personality differences which could be measured by an inventory of this sort.

However, because of the necessity of developing tests for the selection of pilots for military training, after collection of preliminary data (from which the "J" or "Judgment" weights subsequently were developed) the emphasis was shifted. The most readily available criteria of flight proficiency available on military pilots in training (and on student pilots in the Civilian Pilot Training program) was success in the flight training course. Because of this fact the use of the Biographical Inventory has been devoted almost exclusively to the prediction of measures of success in flight training, and the scoring keys for the instrument which have been used extensively have been developed in these terms.

It is quite probable, in this connection, that measures of success in the basic military flight training course (or for that matter in normal civilian training) are not weighted heavily with elements of "safety" and "judgment." In such courses the pilot is not exposed to a wide range of situations requiring the exercise of judgment, in as much as the course is rigidly controlled and much of the flying is done under supervision. More important, under these conditions, the flight instructor or check pilot has little opportunity for assessing the student pilots' judgment except in regard to the specific and limited elements of the flight program.[53] Therefore, an inventory designed primarily to predict "judgment" has been validated, and used primarily to predict an aspect of flight proficiency much more closely associated with "skill."

The nature of what is measured by an inventory of this sort is a function of the nature of the criterion variables in terms of which criterion groups used in determination of the item weights are selected as well as being a function of the nature of the items which constitute the inventory. It cannot be stated, therefore, that the fact that the items originally were phrased with the idea of predicting "judgment" and "safety" explains the relatively low correlations with criterion variables in which elements of judgment and safety do not figure markedly.

It can be pointed out, however, that the philosophy underlying the original development of the inventory was that instruments of this type are particularly useful in predicting those elements of flight performance such as exercise of judgment and safety precautions, which can be considered particularly closely aligned with personality characteristics. The efficiency of the Biographical Inventory in this area has never been

---

[53] As noted previously, this was one of the reasons advanced in explanation of the low correlations found in the early Purdue studies between the "J" weights and success in the controlled Civilian Pilot Training program. Also, of course, the samples on which the "J" weights were derived differed markedly from the sample of student pilots to whose responses on the inventory the weights were applied.

thoroughly explored. Further validation of the type represented by the investigation in which the original [illegible] weights were developed, involving "safe" and "unsafe" civilian pilots, would seem to be suggested.

The Biographical Inventory also would seem to have possible potentialities in the prediction of aspects of military flight performance other than success in flight training, for example, aggressiveness in combat, or suitability for various types of flight duty, i.e., fighter pilot, bomber pilot, etc. In these connections it is perhaps unfortunate that so many of the items in Part B of the Biographical Inventory were eliminated in the course of the several revisions of the instrument. These items which were of the type found in many "personality" inventories, might have been particularly useful in predicting aspects of flight performance less directly connected with flying "skill," or mere ability to handle and to maneuver the plane.

In evaluating the Biographical Inventory in general terms it can be stated that although in research conducted under the auspices of the Committee the correlations between the Biographical Inventory and measures of success in flight training have not been high, nevertheless, the marked utility of the B.I. in the Navy program indicates that the instrument has merit for use in the selection of applicants for flight training, particularly when used as an element in a test battery. It would also seem desirable that future research on this instrument explore its utility in predicting elements of flight performance not predominantely associated with what can most conveniently be termed "skill." It seems probable that the Biographical Inventory may function in predicting differences in flying habits which are related to personality differences, amenable to measurement by an inventory of this type.

SUMMARY

In this report the development of the Biographical Inventory is described and subsequent studies in which it has been used discussed, particularly with reference to findings bearing upon the reliability and validity of the instrument and upon the correlations of the inventory with other tests. Although measures of the reliability of the instrument which have been obtained vary considerably, it is concluded that the reliability of the inventory is sufficiently high to render it serviceable for field use.

In regard to validity of the instrument in most of the studies conducted under the auspices of the Committee on Selection and Training of Aircraft Pilots the correlations between the Biographical Inventory and criterion measures of flight performance have not been high, being in general, slightly lower than the criterion correlations of the Test of Mechanical Comprehension, and on the basis of very limited data, somewhat lower than the criterion correlations of certain psychomotor tests. In these studies criterion correlations of all tests undoubtedly have

been markedly attenuated by the lack of reliability in criterion measures of flight proficiency.

However, although in much of the above research the Biographical Inventory has not predicted measures of success in flight training to a marked degree, the utility of the B.I. in the Navy program indicates that the instrument has merit for use in the selection of applicants for flight training, particularly when used as an element in a test battery. In this connection, the low correlations between the inventory and tests of mental ability indicated in nearly all studies is of significance. It is suggested that, in future research, attention be given to determining the effectiveness of the Biographical Inventory in predicting those aspects of flight proficiency associated with use of judgment and observance of safety precautions.