THE CONSTRUCTION OF HOMOGENEOUS KEYS FOR A BIOGRAPHICAL INVENTORY

Project No. 503–001–0011

Contract No. AF 33(038)–10588

By

PHILIP H. DuBOIS

JANE LOEVINGER

and

GOLDINE C. GLESER

Washington University

PERSONNEL RESEARCH LABORATORY
HUMAN RESOURCES RESEARCH CENTER
AIR TRAINING COMMAND
LACKLAND AIR FORCE BASE
SAN ANTONIO, TEXAS

RESEARCH BULLETIN 52–18
MAY 1952

SUBMITTED BY:
LLOYD G. HUMPHREYS
Director of Research
Personnel Research Laboratory

IMPLICATIONS FOR THE AIR FORCE

"Biographical Inventories" have been used extensively, in conjunction with various tests and procedures, in assessing the aptitudes, interests, and experiences of newly inducted airmen as a basis for effective classification and assignment. The function of a Biographical Inventory is to provide an indication of an airman's aptitudes and interests, as reflected in statements regarding his personal history.

This bulletin describes research on a relatively new technique for scoring or "keying" Biographical Inventories. This procedure, termed "homogeneous keying," involves development of scores in terms of unitary personal characteristics as revealed by replies to items in the Biographical Inventory. There is reason to believe that such homogeneous keys will provide for improved classification, as compared to the more traditional "empirical keys" currently in use. However, further field research to test the utility of these homogeneous keys in the practical situation is planned.

This research bulletin is highly technical and will be of principal interest to research workers.

ARTHUR W. MELTON
DIRECTOR OF OPERATIONS

STEBBINS W. GRIFFITH
COL., USAF
COMMANDING OFFICER

Headquarters, Air Training Command
Human Resources Research Center
Lackland Air Force Base
San Antonio, Texas
21 May 1952

ACKNOWLEDGMENTS

Dr. Lee J. Cronbach was consultant to the project. At the three conferences of Dr. Cronbach and the project staff, many of the ideas presented in this report were clarified.

Mr. Marvin H. Berkeley was the staff member who executed the bulk of the computations. Toward the end of the project his duties were assumed by Mr. Milton A. Whitcomb.

TABLE OF CONTENTS

LIST OF TABLES

THE CONSTRUCTION OF HOMOGENEOUS KEYS FOR A BIOGRAPHICAL INVENTORY

SECTION I

STATEMENT OF PROBLEM

Military experience has shown that biographical inventory data add small but appreciable amounts to the validity coefficients of a number of prediction equations. In broadest outline, the problem to which the present project is addressed is either to increase the validity increment resulting from use of biographical data, or to simplify the work needed to utilize biographical data in new predictions, or both.

Previous use of biographical inventories has depended on the usual "empirical key," in present context better called a "criterion key." A criterion key is constructed separately for each performance for which prediction is needed. Each item is correlated against the criterion and against those items included in the key whose correlation with the criterion exceeds a set standard, such as the .05 level of significance.

There are two practical objections to criterion keys. One is that for a new criterion the procedure must begin again from the beginning. Secondly, experience has shown that the validity of criterion keys shrinks greatly in the cross-validation sample. A more theoretical objection is that criterion keys are factorially complex.

The present study is directed toward finding out whether a particular alternative method of keying is superior. This alternative may be called "homogeneous keying." In this method, the interrelationships of the test items are studied to determine how many homogeneous, relatively independent keys can be formed. It is intended that scores on the homogeneous keys then be utilized in regression equations for various criteria. The advantage of the homogeneous keys lies not only in the fact that it is easier to calculate validity coefficients for about ten variables than to calculate them for about two hundred variables, but also that with the smaller number of variables it is feasible to calculate the complete regression equation. By making use not only of the relationship of each key to the criterion but also of the keys to each other, it is hoped that prediction will be made more efficient. It is also possible, however, that an appreciable number of items with valid variance will be omitted from the homogeneous keys and that, consequently, the homogeneous keys will be less efficient. It is hoped that values derived from homogeneous keys will show less shrinkage on cross-validation.

The inventory used in this study is the Air Force *Biographical Inventory BE601B.* The papers of two samples of airmen were used. Sample A was large enough so that exactly 1000 valid papers were obtained. Sample A was used to construct the homogeneous keys; 850 cases from Sample A for which the needed data were available were used to construct prediction equations for seven criterion keys using the nine homogeneous keys. Sample B consisted of 991 valid papers; it was used to obtain cross-validation data on the homogeneity and independence of the homogeneous keys and to determine the validity of the prediction of the seven criterion keys. It would, of course, have been preferable to have validated the homogeneous keys against criterion scores rather than against criterion keys of the same inventory; however, criterion data were not available on sufficiently large number of cases for that study to be made within the scope of the present project.

As there existed no well-known method for evolving homogeneous independent keys from a large pool of items, it was necessary to evolve simultaneously the method, the theory underlying the method, and the keys. In some instances the method as reported will be more efficient than the one actually used; however, the results do not differ from those that would have been obtained had the more efficient method been used. Some aspects of the method are arbitrary and some are intuitive; it can be doubted, however, whether more rigorous methods are in frequent use.

SECTION II

SOME THEORETICAL CONSIDERATIONS IN MAXIMIZING THE DISCRIMINATING POWER OF A TEST

A. THE CONCEPT OF DISCRIMINATING POWER

In constructing a psychological test one aims either to maximize the validity of the test with an external criterion or to maximize the internal consistency of the test. In recent years there have been a number of different approaches to the problem of constructing homogeneous tests, but the problems raised have not yet been solved to the satisfaction of the majority of psychometricians.

At present a good deal of attention is focused on coefficients to measure the degree of internal consistency of a test, but the "battle of the coefficients" is not entirely realistic. While various writers accuse the coefficients invented by others of being either very good or very bad, the goodness or badness does not inhere in the coefficients. What is needed is a clarification of the purposes of test construction, on the one hand, and the properties of various coefficients, on the other. It will then be apparent which coefficients are most useful for a given purpose.

The essential problem in constructing a homogeneous test may be thought of as maximizing the discriminating power of a test. The discriminating power of a test has three aspects: fineness, probability, and range (4).

How small are the trait differences which can be discriminated by a given test? What is the probability that a given test will correctly discriminate between two people who differ in regard to whatever is measured by the test? Over what range of talent does the test provide discrimination? These three questions are implied in the term discriminating power. They cannot be answered directly, but different proposed coefficients are related more closely to some questions than to others. The questions are far from independent. The fineness of discrimination and the discriminating probability are related negatively or inversely, that is, the finer the discrimination we ask the test to make, the smaller the probability that it will be made correctly. Moreover, the discriminating power of a test in one range of the variable measured may be quite different from that in another range.

The standard deviation of a test is an obvious and important measure of the fineness of discrimination of the test. The standard deviation, however, can be arbitrarily increased by increasing the number of items without limit. As the number of items is increased, the standard deviation will always increase, provided only that the added items are not negatively related to the original ones. Theoretically, increase in the variability of a test under certain circumstances will lessen the value of the test. For a fixed number of items, the maximum value of the standard deviation is obtained when everyone has either the highest or lowest possible score, in which case the discriminating range of the test is reduced to a point. With the magnitude of inter-item covariances found in most psychological tests, however, excessive variance will not be a problem. Excessive variance can be detected when the score distribution departs from normality or rectilinearity in the direction of a U-shaped distribution.

Loevinger's (10) concept of homogeneity is most closely related to the discriminating probability. The discriminating probability of a pair of items with respect to whatever they have in common, which is not necessarily the same as what they have in common with the other items in the test, is given by the coefficient of homogeneity of two items. It can be inferred from Cronbach's (3) Table 7 that the smaller the difference in difficulty between two items, the lower will the value of the coefficient tend to be. The phenomenon in question is essentially the same as the well-known fact that the magnitude of a correlation coefficient depends on the range of talent. This property of the coefficient of homogeneity of two items makes it tend to select as the best pair of items those at opposite extremes of the distribution. In building a test, however, it is preferable to begin with a nucleus of items which discriminate in the middle of the distribution. Items which discriminate one-half of the population from the other half make the maximum number of discriminations and thus make the largest contribution to the discriminating power of the test.

One can arbitrarily increase the homogeneity of a test in Loevinger's sense, or the scalability of a test in Guttman's (7) sense, by decreasing the fineness of the discriminations the test is asked to make, which is, in effect, the course advocated by Guttman. Another way of looking at this property is as follows: Suppose we give

a set of items to a standardization group and determine the interrelations of the items by some appropriate means. We then select the pair of items most closely related to each other. Each item that we add to the test subsequently will lower the coefficient of homogeneity or of scalability of the test. But we know that a test consisting of only two items is not what we wish to achieve; it will be deficient in discriminating range and discriminating fineness.

Cronbach's (3) suggested coefficient, the mean phi coefficient of the items, has the last property also. If we begin by picking the pair of items having the highest phi coefficient, the addition of items to the test will constantly lower the mean phi coefficient. However, Cronbach did not propose this coefficient as a method of test construction.

B. THE CONCEPT OF SATURATION

We may restate the problem of constructing statistically homogeneous tests as follows: Given a finite pool of test items, select a nucleus of two or three items closely related to each other. If we add items in the order of the closeness of their relation to the original nucleus of items, at what point is the discriminating power of the test adversely affected by adding more items? Clearly, we need a measure of discriminating power which will tend to increase at first and will tend to decrease with the addition of the least good items.

Two measures of discriminating power which appear to satisfy these requirements are the Kuder-Richardson (9) formula 20, hereafter called KR 20, and a closely related coefficient which will he called the *saturation coefficient* of the test. The saturation coefficient is defined as the ratio of the sum of all the inter-item covariances to thc total variance of the test. KR 20 is equal to the saturation coefficient times $\frac{n}{n-1}$, where n is the number of items.

The variance of any test may be expanded as a function of the variances and covariances of the items:

$$(1) \quad V_x = \sum_{i=1}^{n} V_i + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} C_{ij}$$

where V_i is the variance of item i, V_x is the variance of the test, and C_{ij} is the covariance of item i with item j. It appears to be a property of a test with high discriminating power that the proportion of the test variance which is due to the covariance of the items is maximized. This proportion is measured by the coefficient of saturation of the test.

The saturation coefficient, S, has now been defined as:

$$(2) \quad S = \frac{2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} C_{ij}}{\sum_{i=1}^{n} V_i + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} C_{ij}} = \frac{V_x - \sum_{i=1}^{n} V_i}{V_x}$$

The term saturation probably dates back to Spearman, as does also the concept of building up tests by building up what they have in common.

> "We can already see, too, that some crude approach towards measuring g can be obtained by the seemingly unscientific course of throwing very miscellaneous tests into a common hotchpot. So doing does not indeed supply an average, or even a representative sample, of the person's abilities . . . . . What the pooling does effect is to make the influences of the many specific factors more or less neutralize each other, so that the eventual result will tend to become an approximate measure of g alone." (11, p. 77)

C. MEASUREMENT OF THE RELATIONSHIP BETWEEN ITEMS

A truly rigorous method of building up homogeneous tests must he based on the relation of every item to every other item. Because of the large number of items in the *Biographical Inventory* studied here, a complete matrix of item interrelations was not obtained. Large parts of this matrix were obtained, however, and a cycling process was followed which was believed to give the same results as would have been obtained had the complete matrix of relationships been obtained.

The problem of what measure to use for the relations between items is thus a central one. The measure used in this study was the inter-item covariance. While the covariance has probably not been used in this fashion in previous test construction projects, it has several advantages.

The covariance has as its upper positive limit the geometric mean of the two item variances, never more than .25. For unrelated items it will equal zero. Iligh values of the covariance will tend to he found among items having close to a .5 split, while items with extreme splits cannot have high covariances.

The effect of using covariances is that the first items included in the test will be those with roughly a .5 split. (Note, however, that with a .4 or a .6 split, the item variance will only be reduced from .25 to .24.) Since such items will in any case provide more discriminations than items with extreme splits, this bias of the covariance appears to be desirable.

The covariance has additive properties that make it useful in test construction. Formulas (1) and (2) show that the variance and saturation of a test can be obtained from the item variances and covariances. The same is true of the correlation between an item and a test and of the correlation between tests. This property makes feasible building tests one item at a time, ascertaining the properties of the test being constructed after each addition of an item.

Finally, the covariance is easier to obtain than the phi coefficient, for example, some problems can be set up so that the covariances are obtained directly from IBM machines.

D. THE PARADOX OF TEST CONSTRUCTION

While KR 20 has recently been defended by Cronbach (3), there are certain outstanding objections to the formula which Cronbach does not answer completely. The same objections hold against the saturation coefficient.

Loevinger's (10) objection to KR 20 is that at the upper extreme values of KR 20, the better tests will have lower values than the poorer tests. The same objection was made to the use of phi coefficients at the upper extreme and, incidentally, holds for the use of covariances.

In regard to the phi coefficient, Cronbach states:

> . . . the phi coefficient which tells when items do and do not duplicate each other is a better index *just because* it does not reach unity for items of unequal difficulty. (3, p. 329)

Here Cronbach neglects the fact that if one uses the phi coefficient for item selection for all ranges of phi, one needs two rules: For lower values of phi, the higher the coefficient, the more will the two items contribute to the discriminating power of the test. But for high values of phi, the lower the coefficient, the more will the two items contribute to the discriminating power of the test.

Similarly, Cronbach shows that the maximum value of KR 20 is not much less than unity for items with a specified distribution of item difficulties, and, that the maximum value will drop for a greater range of item difficulties. If, however, maximizing KR 20 is made a rule for item selection, as Cronbach recommends, there will be a tendency to select items with a narrower range of difficulty, or, in Cronbach's terms, more redundant items. Here Cronbach apparently fails to see the paradox that while maximizing KR 20 will lead to increasing the discriminating power of tests where the item intercorrelations are low, it will lead to decreasing the discriminating power of tests where the item intercorrelations are high.

Tucker states the paradox of test construction as follows:

> Consider the case when all the items in a test are equivalent; that is, when the items all measure the same trait, have equal reliabilities, and are equally difficult. In this case the items are equally intercorrelated with coefficients equal to the item reliabilities . . . . . If the reliability of the items were increased to unity, all correlations between the items would also become unity and a person passing one item would pass all items and another failing one item would fail all items. Thus the only possible scores are a perfect one or one of zero. (12, pp. 1–2)

A similar observation was made by Gulliksen, who stated:

> Such a score distribution is not desirable for obvious reasons, yet current test theory provides no rationale for rejecting such a score distribution. (5, pp. 90–91)

Tucker (12), Brogden (1), and Cronbach (2) have pointed out as a consequence of this paradox that, under certain circumstances, increasing the reliability of a test beyond a certain point will decrease the validity of the test in contrast to the usual belief, embodied in the correction for attenuation, that increasing the reliability of the test always increases its validity.

The paradox of test construction apparently can best be resolved by having two rules for test construction, as indicated above in connection with the phi coefficient, one rule for the construction of tests utilizing items with low intercorrelations and another rule for constructing tests using items with high intercorrelations. Moreover, the papers of Tucker, Brogden, and Cronbach (12, 1, 2) give at least a preliminary indication of how to decide which rule is applicable to a given set of data.

On the basis of the work of Tucker, Brogden, and Cronbach, we may infer that a test will increase in discriminating power the more nearly the items are equivalent, up to a certain point, and decrease in discriminating power the more nearly equivalent the items are after a certain point. Equivalent items are those measuring the

same function and of the same difficulty; further, the best items will be those with difficulty closest to .5. The turning point is a function of the number of items in the test and the intercorrelations of the items. From Tucker's graphs, the mean phi coefficient for a 10–item test should not be higher than about .5, while for a 100–item test, the mean phi coefficient should not be higher than about .25. Substituting these values in Cronbach's (3) equation 44, we obtain the maximum *desirable* value of KR 20 as .91 for a 10–item test and as .97 for a 100–item test. Tests yielding higher values of KR 20 should be treated by alternative test construction methods which do not tend to select equivalent items, such as Guttman's (7) scale analysis or Loevinger's (10) technic of homogeneous tests.

E. HOMOGENEOUS TESTS FOR LOW CORRELATION DATA

Apparently a large part of the test construction efforts of psychologists is expended on data for which the selection of equivalent items will indeed increase the discriminating power of the resultant test. Biographical inventory data certainly falls under this heading.

As Cronbach (3) has shown, one can build up tests of good discriminating power even though the factor which is common to all the items accounts for a small per cent of the variance of each item, say, nine per cent. In constructing such tests, however, it is important not to permit group factors if the final score is to be highly saturated with the common factor. In the present study two devices were adopted, on the basis of intuition and experience, to help avoid group factors within tests.

The first device was that the nucleus of each test was at least three items, rather than the pair of items showing highest covariance. So far as the data permitted, those three items had fairly uniform covariances. It was believed that obvious or concealed "overlapping specifics," to use Spearman's term, might account for a high covariance between a single pair of items, and this initial weighting with specifics might distort the resultant test.

A second device was that at each step of test construction, that is, three-item, four-item, five-item, etc., any item whose inclusion would lower the saturation of the test was permanently excluded from the test. This device was adopted to prevent "functional drift," that is, the inclusion of items measuring function A, then items measuring functions A and B, and finally items measuring function B only. Our experience clearly indicated that functional drift would occur if this rule were not adhered to strictly.

The method of test construction described in this report results in tests for which the covariances fall in a rough hierarchical order. It can easily be seen that the phi coefficients must be equally hierarchical, and each test does thus measure roughly a single factor.

For any four variables (tests or items) the tetrad difference equation is satisfied for their correlations if, and only if, the tetrad difference equation is satisfied for their covariances. The numerators of the correlations are covariances and the denominators of the two pairs of correlations must contain the same four standard deviations:

$$r_{12}\,r_{34} - r_{13}\,r_{24} = \frac{C_{12}\,C_{34}}{\sigma_1\sigma_2\sigma_3\sigma_4} - \frac{C_{13}\,C_{24}}{\sigma_1\sigma_2\sigma_3\sigma_4}.$$

Thus, $r_{12}r_{34} - r_{13}r_{24} = 0$, if, and only if,

$$C_{12}\,C_{34} - C_{13}\,C_{24} = 0.$$

F. MULTIPLE SCORE TESTS

The discriminating power of a multiple score test has one more aspect than the discriminating power of a single test, that is, the degree of independence of the subtests. The discriminating power of a multiple score test is greater, the lower the intercorrelation among its subtests.

Jackson and Ferguson (8) have offered the derivation of KR 20 which appears to be the preferred one at present [Gulliksen, (6)]. They showed that KR 20 is equal to the correlation between two tests which have the same mean inter-item covariance, when the mean item covariance between the two tests is equal to the mean covariance within each. As was to be expected on the basis of this relationship, the saturation coefficient of each test in the present study was found to be the upper limit of its correlations with other tests. When two or more tests were found whose intercorrelations were almost equal to their saturations, these tests could simply have been combined. There was reason to believe, however, that tests with greater discriminating power would result if all the items of the tests involved were

treated as a new pool of items and tests were again constructed beginning with nuclei of three or four items. In the application of the Jackson-Ferguson relationship, the difference between KR 20 and the saturation coefficient is of no importance, as the ratio of the two coefficients is almost one, and in this application attention is not paid to the exact value of the saturation but only to its order of magnitude. This step in the procedure is intended to detect two or more tests of the same or of closely allied functions.

After the most highly saturated tests are constituted from the several pools of items, and after the most highly related tests are reconstituted or combined, there still remain several possibilities for the attenuation of the discriminating power of the tests. (a) An item may have been omitted from a test because it did not fall in the original pool from which that test was drawn. (b) An item may be included in a test even though it is more closely related to another test, or equally closely related to another test.

By dropping some items and adding others it is possible to increase the discriminating power of the multiple score test. Rules for dropping and adding items so as to increase the independence of the tests are based on the table of intercorrelations of the tests, the table of point biserial correlations for every item with every test, and the variances of items and tests. Since the tests are small, it is necessary to correct the point biserial correlation between an item and its own test, for those items included in tests. Considerable judgment is necessary in applying these rules. In general, the only way to reduce the correlation between two tests to zero is to eliminate all of the items from one of the tests. Thus, our aim is to make the intercorrelations low rather than exactly zero. Since the tests change every time an item is dropped or added, the rules must be applied by a successive approximations method. However, by application of these rules, a good many of the items which measure several functions will be dropped from the final multiple score test.

G. COMPARISON WITH FACTOR ANALYSIS

The problem of maximizing the discriminating power of a multiple score test obviously bears some similarity to the problem of factor analysis. In both methods one begins with a large pool of variables, finds the underlying dimensions of the interrelationships of those variables, associates as many as possible of the variables with one of the dimensions, and names the dimensions in terms of the intuitively judged content of its associated variables.

Factor analysis differs from maximizing the discriminating power of a multiple score test in the following ways: (a) Factor analysis has rarely, if ever, been used on pools of much more than 60 variables, while the method of test construction outlined here offered no special difficulties in application to pools of 200 items. (b) Factor analysis seeks to minimize the number of dimensions needed to account for a table of correlations, while the present method seeks to maximize the number of more or less independent tests obtained from a given pool of items. (c) Factor analysis is based on the assumption that each individual's score on each variable is the weighted sum of his scores on the fundamental factors. This assumption seems peculiarly inappropriate when applied to items on which every individual is scored either one or zero. Even as applied to tests, this assumption has been questioned. The present method also begins with a linear equation, equation (1), but it expresses an algebraic identity rather than an assumption. In contrast to factor analysis, which "accounts for" the variance of its variables in terms of per cent weighting of different factors, the present method assigns items to tests on an all-or-none basis.

The idea of assigning as many items as possible to tests which are as independent as possible bears some resemblance to the notion of simple structure, and the cycling process by which tests are refined has some resemblance to rotation of axes. However, assigning variables to hyper-planes, as is done in the search for simple structure, is less stringent than assigning variables to dimensions, as is done in the present method.

The saturation of a test has been defined as the proportion of item covariance to total variance. This quantity, however, is not equal to the proportion of common factor variance according to the usual ideas of factor analysis, for an estimate of the communality of each item should also be included in the common factor variance. KR 20 may be thought of as expressing a lower limit to the proportion of common factor variance.

The proportion of the total variance accounted for by variance in the first factor is a characteristic of the test of great interest. Cronbach (3) states that KR 20 provides an upper limit to this

proportion, but on another page (p. 322) he interprets KR 20 as directly equal to the proportion of first factor variance in a particular example. This interpretation is justified only when there is substantial evidence of hierarchieal order within the test. Some evidence of hierarchy does exist for tests constructed by the method described here.

Since KR 20 is related only approximately to its factorial interpretations, its interpretability does not appear to be a reason for preferring it to the saturation coefficient.

SECTION III

MAXIMIZING THE SATURATION OF A TEST

The procedure followed in maximizing the saturation of each key of the *Biographical Inventory* was as follows: From a given matrix of inter-item covariances, the highest triplet was chosen. That is, the nucleus of the key was the set of three items with highest covariances *inter se*. Those three items comprise a test. All items are discarded which would lower the saturation of that three-item test. The one item is added which would maximize the saturation of the resultant four-item test. Then all items are discarded which would lower the saturation of that four-item test, and the one is added which would maximize the saturation of the resultant five-item test, and so on. The process terminates when all items are either included in the test or excluded from the pool. Items are included in some of the (Cycle 1) keys of this study which raised the saturation only in the third decimal place. A more rigorous rule, resulting in purer tests, would have been to include only those items raising the saturation in the second decimal place.

The first step in evolving a computational procedure is to note that in order to maximize the saturation one need only maximize a simpler quantity:

$$(3)\qquad {}_nW_t = \frac{\sum_{i=1}^{n-1}\ \sum_{j=i+1}^{n} C_{ij}}{\sum_{i=1}^{n} V_i},$$

where C_{ij} refers to the covariance of items i and j, V_i refers to the variance of item i, the subscript t on the ratio W means that it is a property of the test, and the prescript n refers to the number of items in the test. The quantity ${}_nW_t$ changes every time an item is added to the test.

The proof that maximizing ${}_nW_t$ will maximize the saturation is simple. Note first that the saturation is a quantity of the form $\frac{2C}{V+2C}$, where the capitals without subscripts are used to designate the sums rather than the elements of the sum. To maximize the saturation one needs only to minimize its reciprocal, $\frac{V+2C}{2C} = \frac{V}{2C}+1$. As constants may be disregarded, one needs to minimize $\frac{V}{C}$, or maximize $\frac{C}{V}$. The next step is to find a criterion for the exclusion of items. Let us define a ratio ${}_nW_k$ characterizing each item k not included in the test:

$$(4)\qquad {}_nW_k = \frac{\sum_{i=1}^{n} C_{ik}}{V_k}$$

where the subscript k indicates that the W is a property of the item k, the prescript n means that there are n items in the test, k not being one of the first n items in the test. As before, C_{ik} is the covariance of items i and k, and V_k is the variance of the item k. It can be shown that an item k will not lower the saturation of the test if

$$(5)\qquad {}_nW_k \geq {}_nW_t.$$

The proof that the inequality (5) establishes that item k will not lower the saturation is as follows. Assume that item k will not lower the W_t ratio. This condition may be expressed:

$${}_{n+1}W_t \geq {}_nW_t.$$

Substituting from equation (3):

$$\frac{\sum_i C_{ik} + \sum_i \sum_j C_{ij}}{V_k + \sum_i V_i} \geq \frac{\sum_i \sum_j C_{ij}}{\sum_i V_i}.$$

Since all variances are positive, we may multiply by the denominators without changing the sign of the inequality.

$$\sum_i V_i \left(\sum\sum_{ij} C_{ij} + \sum_i C_{ik}\right) \geq \sum\sum_{ij} C_{ij} \left(V_k + \sum_i V_i\right).$$

Cancelling like terms and dividing again by the variance terms yields

$$\frac{\sum_i C_{ik}}{V_k} \geq \frac{\sum\sum_{ij} C_{ij}}{\sum_i V_i}.$$

Since every step in the proof is reversible, the

inequality expressed in formula (5) is established. The same proof may be used to show that item k will lower the saturation if ${}_nW_k < {}_nW_t$.

Apparently a criterion corresponding to that of formula (5) for whether an item will lower KR 20 will be more complicated.

Worksheets for constructing tests by the present method are shown in Tables 1 and 2, which must be constructed simultaneously. The right side of Table 1 consists of a table of covariances for items included in the test.

After the original nucleus of items is chosen, the first three covariances are entered on the right side of Table 1. The sum of the first three covariances is entered in the (3) (3) cell of the principal diagonal. The variances of the first three items are entered in the first column to the left of the vertical item identification, and the sum of the first three variances is entered in the next column leftward. The first test ratio is entered in the leftmost column; it is equal to the ratio of the sum of the first three covariances to the sum of the first three variances.

At this stage it is convenient to have Table 2 drawn up but no entries made in it. For each item in turn the quantity ${}_3W_i$ is now computed. If the ratio for the item exceeds the ratio for the test, then the identifying symbol of the item is entered in the first row, its variance is entered in the same column, second row, and the sum of its covariances with the first three items is entered in the same column, third row. This step is completed for the entire original matrix of items. Most of the items will be rejected at this step and thus will not appear in either table.

The next step is to compute a trial ${}_4W_t$ for each item in Table 2. The trial ${}_4W_t$ is equal to the sum of covariances of the test plus the sum of covariances of the item, divided by the sum of variances for the test plus the item variance. The values for the test are found in Table 1, the corresponding values for the item are found in the appropriate column of Table 2. With some practice it is not necessary to compute all trial values. Items with low variance will not be selected when there are items with high variances in the pool, for example. Usually the item with the greatest difference between its variance and the sum of its covariances will be selected.

The item which has the highest trial ${}_4W_t$ is selected as the fourth test item. Its covariances with the three items already in the test are entered in the right side of Table 1, and its variance is entered in the column of Table 1 labelled V_i. The three covariances just entered in the table are now added to the previous total, found in cell (3) (3), and the new total is entered in cell (4) (4). The new sum of variance is obtained by adding the new variance to the previous sum of variances. The new test ratio, ${}_4W_t$, is obtained by dividing the sum of covariances by the sum of variances. The value obtained should check

TABLE 2

POOL OF ITEMS: A SAMPLE TAB LE

	Item	59	69	70	95	96	124
	V_i	.2485	.1957	.2481	.2402	.2491	.1124
	${}_3\Sigma C$	.0882	.0645	.0738	.1174	.0767	.1122
	${}_3W_i$	.355	.330	.297	.489	.308	.998
Trial	${}_4W_t$	.328	.321	out	.362	out	.411
	${}_4\Sigma C$	.0987	.0721	---	.1371	---	in
	${}_4W_i$	.397	.368	---	.571	---	---
		out	out	---	in	---	---

TABLE 1

SYNTHESIS OF TEST STATISTICS: A SAMPLE TABLE

${}_nW_t$	ΣV_i	V_i	Item	Item Covariances			
				117a	110b	124a	95a
		.2447	109b	.0850	.0775	.0309	.0482
		.2483	117a		.0642	.0418	.0320
.319	.7113	.2183	110b	${}_3\Sigma\Sigma C =$	.2267	.0395	.0371
.411	.8237	.1124	124a		${}_4\Sigma\Sigma C =$	.3389	.0197
.447	1.0639	.2402	95a			${}_5\Sigma\Sigma C =$	.4759

exactly with the corresponding value in the "Trial $_4W_t$" row of Table 2. It will be convenient to draw a heavy line down the column of Table 2 corresponding to the item selected for the test. This line indicates that the item is no longer in the pool and facilitates inspection of Table 2 during later computations.

For each item a new sum of covariances is obtained by adding its covariance with the fourth item to its previous sum of covariances. The values are entered in the row of Table 2 labelled $_4\Sigma C$. The sum of covariances for each item is divided by its variance. These ratios need not be recorded, but, for those items where the ratio is less than the test ratio, an indication must be made that the item no longer is in the pool. This indication may again be a heavy line down the column. For those items remaining in the pool, a trial $_5W_t$ is computed, and so on.

At any stage the parameters of the test may be ascertained. The variance of the test will equal the sum of the item variances plus twice the sum of all the covariances. The saturation will equal twice the sum of the covariances divided by the test variance. KR 20 will equal $\frac{n}{n-1}$ times the saturation. The mean of the test will equal the sum of the proportions answering plus to each item; these values are not included in Tables 1 or 2 but must be known in order to compute the item variances. It is not necessary to compute the test parameters as each item is added, but, if it is desired to do so, another table giving the above values can easily be drawn up.

SECTION IV

A METHOD FOR MAXIMIZING THE DISCRIMINATING POWER OF A MULTIPLE SCORE TEST

The problem of this research was to resolve a large pool of biographical inventory items into a set of homogeneous and relatively independent subtests or keys. Methodologically, the problem would have been the same if the basic data had been items of many other types of tests, such as interest tests or multiphasic personality tests. The method evolved can be used for the discovery of traits or of types of people; there appear to be no assumptions which limit it in this respect.

The method has been worked out for the following conditions: (a) Items were either given as dichotomous or reduced to dichotomous form. (b) There were not many items with very high intercorrelations. (c) IBM equipment was available for the analysis of the data. (d) Since we had only a rough idea of the sampling errors involved, a large number of cases was required. The use of exactly 1000 cases in the major part of the study saved many hours of labor, since all divisions by N were accomplished by shifting the decimal place.

With some modifications, which have not yet been worked out, the method could be applied to multiple choice items or to items showing high intercorrelations. Probably no test construction technique can usefully be applied to small numbers of cases; however, cross-validation data from the present study appear to indicate that useful results might be obtained with as few as 300 cases.

Essentially the method is based on the matrix of the covariances of every item with every other one. With fewer than 100 or so items, the covariances of all the items can actually be determined. With pools of items in the neighborhood of 200, there are complications in obtaining and in handling the complete matrix of covariances. The present method is adapted to the latter case, but it can easily be simplified to cover the former.

The first step involves a careful reading of the test and the formulation of hypotheses as to possible interrelations of the items. Items are then assigned to a priori matrices according to these hypotheses and apparent similarity of content. It is convenient also to consider the capacity of the IBM machine being used in constructing the a priori matrices. With 1000 cases, 19 is a convenient upper limit to the number of items in a matrix. With 19 items, a single run of the cards through an 80 counter tabulator will yield all the item counts and a card count, and the results can be printed with a space between every pair of numbers. Where it is not possible to limit the matrix to 19 items without doing violence to the content, then 38 should be taken as the upper limit of the number of items. Two or more a priori groupings of items may be included on a single IBM card if they total no more than 19 items. Items not included in any a priori grouping may be assigned arbitrarily wherever there is space on the cards, to bring the total number of items up to 19.

Cycle I keys are evolved from the a priori matrices by the method described in the previous chapter. After one key is constructed from a matrix, the entire original matrix is utilized in constructing further keys. It was thought desirable at first to exclude those items in the first key from consideration for later keys, but this course is probably disadvantageous. An item which is drawn into the first key as one of the last items may more properly appear as one of the first items of a second key. It would then probably belong with the second key. Items closely related to both keys may often best be omitted from both, since they tend to raise the correlation between the two keys.

All items which are not included in any key are duplicated onto a new card, termed the residual matrix. The residual matrix is treated the same way that the a priori matrices are treated, that is, the covariances of all items are obtained and the total matrix examined for new keys. The keys derived from the a priori matrices plus those derived from the residual matrix now constitute Cycle I keys. Cycle I keys are scored and correlated; if an IBM automatic multiplier is available, the scoring can be done by cross-footing. The matrix of intercorrelations of Cycle I keys is examined for high values, say, values above .25 or .3. These are the correlations which must be reduced in order to have relatively independent tests, and so far as possible this reduction must take place without impairing the saturation of the tests.

If there are two or more keys which have correlations *inter se* approaching in magnitude their saturations, it is known from the Jackson-Ferguson principle that the covariances of items in one key with the items in the other key or keys are about the same as the covariances of items within each key. All of the items in such a closely related group of keys are duplicated onto a new card from which Cycle IA keys are constructed to replace the corresponding Cycle I keys. There may be two or more such groups of closely related keys. Each group of keys is, of course, treated separately. Cycle IA keys are constructed by the same method used for Cycle I tests. Cycle IA keys are now scored and correlated with each other and with those of the Cycle I keys which are retained without change.

The next step is to obtain the point biserial correlation of every key, i.e., the Cycle IA keys plus the Cycle I keys that were not replaced in Cycle IA, with every item in the original pool. These correlations comprise a matrix with one column for each key and one row for each item in the pool, though of course the data need not be printed in that form. In general, it is necessary to apply a correction to the point biserial between the item and its own key to compensate for the spurious correlation. In practice, for many items the outcome will by inspection either be so high or so low that the actual computations need not be made. The formula for this correction is as follows:

$$(6) \qquad r_{i\,(T-i)} = \frac{r_{iT}\sigma_T - \sigma_i}{\sqrt{\sigma_T^2 - 2r_{iT}\sigma_i\sigma_T + \sigma_i^2}}$$

where $r_{i(T-i)}$ is the corrected point biserial, r_{iT} is the uncorrected point biserial, and σ_i and σ_T represent standard deviations of item and key. A useful approximation is given by:

$$(7) \qquad r_{i(T-I)} \cong r_{iT} - \frac{\sigma_i}{\sigma_T}$$

The matrix of point biserials is utilized to drop items from, or add items to, keys, primarily to lower the correlations between keys but, in some cases, also to raise the saturation. There are three major considerations in examining this matrix. The first is that every item should have its highest correlation with its own key. If an item has a higher correlation with another key than with its own, it should be dropped from the one it is in. It may be added to the key with which it has a higher correlation but usually not. Items with fairly equal correlation with two or more keys are often best omitted entirely, since they are the items which raise the correlations between keys. Occasionally one will find that key A and key B will be positively correlated but that item i will enter A in a positive sense and enter B in a negative sense. In this case, inclusion of the item in both keys acts to lower the correlation between them.

The second consideration is that some items not included in any Cycle I key will have a high correlation with just one of those keys. This will occur only when the item was not included in the matrix from which that test was drawn. Care must be taken not to add items to a key if those items will raise correlations which are already high.

The third consideration is to lower high correlations between keys. For every pair of keys having a high correlation, say, over .25, every item in both tests should be examined to see if it has a fairly high correlation with the key in which it is not included. Of course any items included in both keys would be dropped from one or both. In dropping items care should be taken not to deplete any test to the point where its saturation falls below .4 or at least .35.

When the complete matrix of covariances is available, the correlations between any key and any other or between a key and any item can be quickly recomputed after each deletion or addition of an item to the key. When the complete matrix of covariances is not available, the most practical procedure is to make the few changes for each test which are most clearly indicated. After such changes, the new tests are called the Cycle II tests. The Cycle II tests are scored, correlated, and the biserial correlation of each test with each item is again obtained. The same considerations are applied to obtain Cycle III tests, and so on. The process terminates automatically when there are no further changes.

The following formulas are useful in carrying out the cycling process. If item i is not included in either test T_1 or test T_2, then adding i to T_1 will not raise the correlation between T_1 and T_2 if

$$(8) \qquad \frac{r_{iT_2}}{r_{iT_1} + \frac{\sigma_i}{2\sigma_{T_1}}} \leq r_{T_1T_2}.$$

If item i is included in T_1, then the correlation between T_1 and T_2 will be lowered by dropping i from T_1 if

$$(9) \qquad \frac{r_{iT_2}}{r_{iT_1} - \frac{\sigma_i}{2\sigma_{T_1}}} \geq r_{T_1T_2}.$$

The test ratio of formula (4) can be obtained from the point biserial correlation by means of the following formula:

$$(10) \qquad {}_nW_i = r_{iT} \frac{\sigma_T}{\sigma_i}.$$

Unity must be subtracted from the right hand side if the item i is included in the test T. This formula enables one to determine whether a given item will lower the saturation of a key when the item was not included in the matrix from which the key was drawn.

SECTION V

CONSTRUCTION AND CROSS-VALIDATION OF HOMOGENEOUS KEYS FOR THE BIOGRAPHICAL INVENTORY BE601B

The data analyzed to derive homogeneous keys for the *Biographical Inventory BE601B* were the answer papers of 1000 airmen from consecutive flights. In order to simplify statistical work, only completely and correctly marked papers were used. Preliminary scanning eliminated about 40 per cent of the papers, which had no marks for items which required at least one mark, or several marks for items which prohibited more than one mark. The first 1000 complete cases were termed Sample A.

Of the 1000 cases in Sample A, scores on the seven Air Force empirical, or criterion, keys were available for 850 cases. These 850 cases were used for the correlation of the final homogeneous keys with criterion keys.

A second sample, Sample B, used for cross-validation purposes, was constituted in the same fashion as Sample A. It contained 991 cases, for all of whom there were scores on the nine homogeneous keys derived in this study and the seven Air Force criterion keys.

In the description of the test construction process which follows, all data not otherwise specified were obtained from the 1000 cases of Sample A.

The *Biographical Inventory BE601B* contains a total of 125 items; however, many of these are five choice items. Of the five choice items, some were worded so that no one could correctly mark two of the choices, while for others several marks were permissible. Even where only one choice was permitted, the five choices usually did not lie along a single continuum but were often what is known as "double-barrelled" items.

The first step in the construction of homogeneous subtests or keys was to obtain item counts for each alternative of each item. These item counts were utilized, together with a careful reading of the items, to dichotomize a number

of the multiple choice items. So far as possible, where no violence was done to the sense of the question, the items were dichotomized so as to produce a split at the median. For many questions there was no adequate way of doing justice to the content except to record each alternative as a separate item. A total of 175 items was punched on three master IBM decks, identified as decks A, B, and C.

Next, the items were assigned to groups according to apparent similarity of content. There were 16 such groups of items. Some of these groups were small and were therefore combined on a single IBM card. The a priori matrices were punched on 10 decks of cards, labelled decks D through M. The tentative names given the a priori matrices are indicated in Table 3. Characteristics of Cycle I keys are shown in Table 4. It will be noted that each key is denoted by a symbol which indicates the matrix of items from which that test was originally drawn. The subscript denotes whether the key was the first or

TABLE 3

A PRIORI MATRICES

Deck	A Priori Categories	Topics Included
D	Mechanical	Automobile, airplane, ordinance, blueprints.
E	Shop Work Precision Work	Carpentry, metal work, odd jobs. Watches, instruments, model planes.
F	Electrical	Radio except operator, electrical, radar.
G	Scientific (physical) Scientific (biological)	Weather, radar fundamentals, mathematics. Medical, animals, agriculture.
H	Clerical-Computational	Cryptography, business administration, commercial, puzzles, office machines.
I	Security-Routine	Radio operator, photography technician, parachutes, collecting, personal service.
J	Ambition Social	Earning, profit, contests, boss, college. Sports, games.
K	Responsibility	Inspector, control tower, traffic, use of compass, delegating responsibility.
L	Selling Teaching-Persuasion Social Service	 Public speaking, acting, technical instructor. Medical corps, psychology, scouting, social studies.
M	Artistic Literary	Drafting, photography. Foreign language, journalism.

TABLE 4

CHARACTERISTICS OF CYCLE I KEYS

Key	Symbol	No. of Items	Mean	Variance	Final n Weight	Saturation
Mechanical	D	15	6.8	11.9	1.41	.74
Woodshop-Precision	E	10	4.3	5.37	.74	.60
Radio Technician	F	11	4.3	7.95	1.36	.73
Scientific-Education	G	10	3.5	4.18	.53	.52
Clerical-Computational	H	11	2.4	4.94	.96	.66
Security-Routine	I	6	2.3	1.88	.27	.35
Response Set (?)	J_1	6	3.7	2.49	.49	.50
Ambition	J_2	9	4.0	3.95	.49	.50
Responsibility	K_1	6	3.1	1.95	.21	.30
Supervisory	K_2	5	1.9	2.02	.45	.47
Extroversion (?)	L	17	5.4	7.84	.91	.65
Drafting	M_1	6	2.6	2.35	.37	.42
Literary	M_2	9	2.4	2.75	.44	.47
White Collar	N_1	9	4.7	3.23	.38	.43
Deciphering	N_2	6	2.2	2.02	.34	.40

second drawn from the matrix. The first key drawn from a matrix usually has the most tightly related nucleus of items.

No more than 25 items were included in a single a priori matrix. This number was arrived at because it was about the maximum number which permitted item counts to be obtained for all items with a single wiring of the IBM tabulator, which contained 80 counters. Later it was believed preferable to limit the number of items in the original matrices to 19, so that a blank space would intervene between any two item counts as they were printed by the tabulator, and a card count could be obtained simultaneously. In either case, sufficient space was left on the cards to permit a number of later operations to be carried on with the same decks.

Cycle I keys were derived from the covariances of items in the a priori matrices by the method described in Section III. The items not included in any Cycle I key were reproduced onto a new deck, deck N, the residual matrix. From this matrix two additional keys were obtained. These keys are also included as Cycle I keys in Tables 4 and 5.

Cycle I keys were scored by means of cross-footing on an IBM multiplier. These scores were then correlated to produce the coefficients recorded in Table 5. Inspection of Table 5 showed two pools of keys with correlations *inter se* approaching their saturations.

One pool consisted of the keys labelled Woodshop-Precision Work, Drafting, and Deciphering. The Deciphering key contained items referring to reading of blueprints, among others. The items from these three tests were treated as a new matrix. This matrix yielded a single key, Handyman, having a higher saturation than any of the three keys from which it was formed.

The second pool consisted of keys labelled Scientific Education, Ambition, Extroversion (?), and Literary. The matrix formed from these items yielded two keys which cut across the previous ones, although the most nearly appropriate names were those of two of the previous keys, Extroversion (?) and Education. The (?) following Extroversion is meant to indicate that the content of the items is not well summarized by the term. No single more appropriate term suggested itself, however. Characteristics of the Cycle IA keys are shown in Table 6.

Cycle IA keys were scored and correlated with each other and with those keys retained unchanged from Cycle I. The correlations are displayed in Table 7. The highest correlation was one of .51 between Extroversion (?) and Education. There

TABLE 5

INTERCORRELATIONS OF CYCLE I KEYS*
(Diagonal Entries Are Saturation Coefficients)

Key	Symbol	D	E	F	G	H	I	J_1	J_2	K_1	K_2	L	M_1	M_2	N_1	N_2
Mechanical	D	74														
Woodshop	E	36	60													
Radio	F	15	28	73												
Education	G	-16	12	18	52											
Clerical	H	-15	-14	09	35	66										
Security-Routine	I	00	-12	-10	-18	00	35									
Response Set (?)	J_1	05	14	10	16	11	-02	50								
Ambition	J_2	-08	12	14	34	22	-14	29	50							
Responsibility	K_1	15	30	31	31	07	-17	10	27	30						
Supervisory	K_2	-04	03	05	23	17	-19	12	30	16	47					
Extroversion (?)	L	-06	10	13	61	37	-14	25	53	35	31	65				
Drafting	M_1	09	49	16	18	-14	-06	07	17	17	05	12	42			
Literary	M_2	-15	02	06	35	23	-17	30	22	18	16	42	05	47		
White Collar	N_1	04	10	17	18	18	04	06	10	13	15	20	10	12	43	
Deciphering	N_2	19	43	30	25	05	-10	12	19	35	11	28	25	16	18	40

*Decimal points have been omitted.

were 13 correlations exceeding .3 in absolute value.

The point biserial correlations between the 11 Cycle IA keys and all of the items, including those not incorporated in any key, were computed next. This operation was set up to be performed on the IBM Automatic Multiplier. Items were judged for exclusion from or inclusion in keys on the basis of the considerations discussed in Section IV. These criteria led to dropping all of the items in the key labelled Responsibility; therefore, this key was dropped. A number of items were dropped from other keys and a few added.

The results of these changes gave Cycle II keys, whose characteristics are shown in Table 8. Intercorrelations of Cycle II keys are shown in Table 9. In this table only three correlations exceed .3. The point biserial correlations were again computed between every item and the Cycle II keys. All items in the White Collar key were more closely related to other keys than to their own; therefore this key was dropped. A few items were added to or dropped from other keys.

The tests resulting from these changes comprised the Cycle III keys. As there were only a few changes in Cycle III, no further cycles were made. It was judged that few if any changes would be made in further cycles and that in any case further statistical refinement was less important than utilization of the findings thus far for the improvement of the content of the test.

At this point it may be well to characterize the content of the subtests or keys as they emerged in Cycle III. The Mechanical key contains items referring to experience with and interest in mechanical things; automotive repairs are an important aspect.

The Radio key contains items referring primarily to interest in but also to experience with various aspects of radio work.

TABLE 6

CHARACTERISTICS OF CYCLE IA KEYS

Key	Symbol	Keys Replaced*	Number of Items	Mean	Variance	Final n Weight	Saturation
Handyman	O	(E, M_1, N_2)	17	7.4	13.0	1.23	.71
Extroversion (?)	P_1	(G, J_2, L, M_2)	16	7.1	11.3	1.19	.71
Education	P_2	(G, J_2, L, M_2)	15	4.6	7.37	.83	.63

*Keys not indicated in "Keys Replaced" column are retained unchanged in Cycle IA.

TABLE 7

INTERCORRELATIONS OF CYCLE IA KEYS*

(Diagonal Entries Are Saturation Coefficients)

Key	Symbol	D	F	H	I	J_1	K_1	K_2	N_1	O	P_1	P_2
Mechanical	D	74										
Radio	F	11	73									
Clerical	H	-33	09	66								
Security-Routine	I	-02	-10	00	35							
Response Set (?)	J_1	00	10	11	-02	50						
Responsibility	K_1	13	31	07	-17	10	30					
Supervisory	K_2	-09	05	17	-19	12	16	47				
White Collar	N_1	-21	20	32	04	06	12	21	43			
Handyman	O	37	31	-18	-12	12	31	05	07	71		
Extroversion (?)	P_1	-17	12	34	-14	32	34	35	24	14	71	
Education	P_2	-26	17	34	-24	21	29	27	33	13	51	63

*Decimal points have been omitted.

TABLE 8

CHARACTERISTICS OF CYCLE II KEYS

Key	Symbol	No. of Items	Mean	Variance	n Weight	Saturation
Mechanical	D	9	3.7	6.73	1.16	.70
Radio	F	14	5.4	13.3	1.79	.78
Clerical	H	8	1.5	3.00	.77	.61
Security-Routine	I	6	2.3	1.88	.27	.35
Response Set (?)	J_1	8	4.5	3.89	.71	.59
Supervisory	K_2	5	1.9	2.02	.45	.47
White Collar	N_1	7	4.3	2.30	.30	.38
Handyman	O	15	6.7	11.2	1.14	.70
Extroversion (?)	P_1	13	6.4	8.39	.99	.66
Education	P_2	9	2.7	3.70	.71	.59

TABLE 9

INTERCORRELATIONS OF CYCLE II KEYS*

(Diagonal Entries Are Saturations)

N=1000

Key	Symbol	D	F	H	I	J_1	K_2	N_1	O	P_1	P_2
Mechanical	D	70									
Radio	F	12	78								
Clerical	H	-12	11	61							
Security-Routine	I	00	-10	00	35						
Response Set (?)	J_1	05	13	08	01	59					
Supervisory	K_2	-07	06	14	-19	08	47				
White Collar	N_1	-09	17	15	07	01	51	38			
Handyman	O	31	25	-09	-11	15	06	09	70		
Extroversion (?)	P_1	-13	16	27	-13	21	31	15	17	66	
Education	P_2	-19	20	21	-23	15	22	24	11	40	59

*Decimal points have been omitted.

The Clerical key contains items referring to interest in and experience in clerical work.

The Security-Routine key contains items concerning preference for secure and routine jobs.

The key now labelled Response Set offered the greatest difficulty in interpretation. Hypotheses such as introversion and conventionality were entertained as to the common content. Since all the items occur in a long check list on page 7 of the test booklet, the hypothesis of a response set is most likely. There may also be other psychological factors entering into the relationships of some pairs of items. From the point of view of homogeneity, this key is one of the stronger ones; however, in the regression equations derived from Table 13, discussed later, it had no high weights.

The Supervisory key contains items referring to a preference for supervisory work.

The Handyman key, in addition to those items suggested by its title, contains items referring to experience and interest in drafting and in reading blueprints.

The Extroversion (?) key also offers difficulties of interpretation. It contains items referring to leadership, public speaking, literary or verbal

interests, and interest in people. In the case of this key, one has the impression that the statistical methods used were not sufficiently rigorous to prevent "functional drift." Very likely, by adding new items and re-analyzing the test, one could break down this subtest into at least two more or less independent ones.

The Education key is weighted with items referring to scientific education as well as education in general.

Table 10 shows the characteristics of the Cycle III keys for Sample A and for Sample B. Sample A was, of course, the sample whose data were used in assigning items to keys. No data from Sample B were used in constructing or refining keys. Therefore, one would expect saturations to be slightly lower in Sample B. It will be noted from Table 10 that the "shrinkage" of the saturations in Sample B is not great.

Tables 11 and 12 show the intercorrelations of the Cycle III keys for Samples A and B respectively. In each of these tables there are just two correlations above .3. It is not obvious whether one would expect the correlations to rise or fall in the cross-validation sample. Because just those items were removed from the various keys which raised the correlations between keys, one might expect the correlations to rise slightly for

TABLE 10

CHARACTERISTICS OF CYCLE III KEYS

Key	Symbol	Number of Items	Standardization, Sample A (N=1000) Mean	SD	n^W_t	Saturation	Cross-Validation, Sample B (N=991) Mean	SD	Saturation
Mechanical	D	12	5.7	3.16	1.55	.76	5.7	3.13	.72
Radio	F	14	5.4	3.65	1.79	.78	5.4	3.46	.76
Clerical	H	9	1.8	1.95	.87	.64	1.7	1.84	.60
Security-Routine	I	7	2.9	1.57	.36	.42	3.0	1.54	.37
Response Set (?)	J_1	10	5.8	2.32	1.06	.68	5.8	2.43	.63
Supervisory	K_2	5	1.9	1.42	.45	.47	1.9	1.38	.45
Handyman	O	13	6.0	2.78	1.07	.68	6.1	3.07	.69
Extroversion (?)	P_1	14	6.6	3.05	1.06	.68	6.5	3.08	.68
Education	P_2	7	1.9	1.60	.50	.50	1.8	1.53	.48

TABLE 11

INTERCORRELATIONS OF CYCLE III KEYS, SAMPLE A*

(Diagonal Entries Are Saturations)

N = 1000

Key	Symbol	D	F	H	I	J_1	K_2	O	P_1	P_2
Mechanical	D	76								
Radio	F	09	78							
Clerical	H	-22	08	64						
Security-Routine	I	03	-09	04	42					
Response Set (?)	J_1	05	13	08	-01	68				
Supervisory	K_2	-11	06	14	-18	09	47			
Handyman	O	25	27	-19	-12	17	07	68		
Extroversion (?)	P_1	-15	16	27	-11	23	31	20	68	
Education	P_2	-24	20	19	-19	09	20	15	34	50

*Decimal points have been omitted.

TABLE 12

INTERCORRELATIONS OF CYCLE III KEYS, SAMPLE B*

(Diagonal Entries Are Saturations)

N=991

Key	Symbol	D	F	H	I	J_1	K_o	O	P_1	P_2
Mechanical	D	72								
Radio	F	15	76							
Clerical	H	-22	04	60						
Security-Routine	I	04	-02	01	37					
Response Set (?)	J_1	03	05	16	-08	63				
Supervisory	K_2	-11	-02	13	-15	11	45			
Handyman	O	24	17	-15	-03	14	02	69		
Extroversion (?)	P_1	-08	12	37	-10	21	27	26	68	
Education	P_2	-17	16	22	-17	14	20	18	31	48

*Decimal points have been omitted.

TABLE 13

CORRELATIONS* OF CYCLE III KEYS WITH CRITERION KEYS, SAMPLE A

N=850

		Criterion Keys						
Key	Symbol	Clerical	Mechanical	Craftsman	Equipment Operator	Radio Operator	Services	Electronics
Mechanical	D	-49	13	-12	53	-24	-23	-17
Radio	F	20	27	02	-24	44	-56	31
Clerical	H	52	-15	-28	-28	32	11	03
Security-Routine	I	-20	-32	-11	08	-19	09	-27
Response Set (?)	J_1	19	12	-02	-15	13	-05	13
Supervisory	K_2	32	22	04	-27	26	38	32
Handyman	O	03	61	51	-08	36	-37	49
Extroversion (?)	P_1	52	31	03	-39	52	-01	43
Education	P_2	63	44	18	-45	56	-05	56
Multiple R		85	77	63	71	79	75	77
Cross-Validation R		83	76	57	69	81	74	77

*Decimal points have been omitted.

Sample B. But the saturations fall slightly, and as the correlation between variables is to a considerable extent a function of their saturations, one might expect the correlations to fall slightly. The mean of the absolute values of the correlations in Table 11 is .15; the corresponding value for Table 12 is .14. These values appear to indicate a satisfactory degree of independence for the homogeneous keys. By contrast, the mean of the absolute values of the intercorrelations of the criterion keys is .39 for 850 cases of Sample A. It should be remembered, however, that the correlations between criterion keys are somewhat spurious, since there are a number of instances of the same item entering two or more keys.

The most important question to be asked concerning the homogeneous keys is how well they aid in prediction. An adequate answer to this question is beyond the scope of this project. As a preliminary indication of the predictive validity of the homogeneous keys, the multiple correlation of the nine Cycle III keys with each of the seven criterion keys was computed. These multiple correlations are shown in Table 13, together with the first-order correlations of homogeneous with criterion keys. The intercorrelations of

TABLE 14

INTERCORRELATIONS OF CRITERION KEYS, SAMPLE A*

N=850

Criterion Key	Clerical	Mechanical	Craftsman	Equipment Operator	Radio Operator	Services
Mechanical	34					
Craftsman	14	52				
Equipment Operator	-64	-25	-13			
Radio Operator	74	56	29	-51		
Services	06	-24	-06	-13	29	
Electronics	57	84	46	-65	68	-09

*Decimal points have been omitted.

criterion keys are presented in Table 14. The multiple R's range from .63 to .85.

The row of Table 13 labelled "Cross-Validation R" was computed from data of Sample B as follows: The 991 papers of Sample B were scored according to the nine homogeneous keys worked out on Sample A. They had previously been scored on the seven Air Force criterion keys. The scores on the homogeneous keys were utilized in prediction equations, with weights taken from the multiple regression equations of Sample A. The cross-validation R's were the correlations between predicted and actual values of the criterion keys. Since the weights were determined so as to maximize the multiple R for Sample A, one would expect some shrinkage of the cross-validation R as compared to the multiple R in Sample A. This shrinkage is observed but is small.

It would be unwise to attach great significance to the correlations reported in Table 13. Each of the criterion keys was based on the scoring of a considerable number of items, and many or most of these items were also scored in the homogeneous keys, sometimes in the same sense, sometimes in the opposite sense. Thus, the multiple R's are spurious to an unknown degree.

One may conclude that the construction of homogeneous independent keys by the method of the previous chapters is feasible and results in subtests of psychological interest.

SECTION VI
SUMMARY

1. A method of constructing homogeneous independent keys for a biographical inventory has been presented. The method is based on analysis of the inter-item covariances, and it appears to maximize the discriminating power of the resultant multiple score test.

2. Each key is constructed by adding items one at a time to a nucleus of three items. Each key is constructed so as to maximize the saturation with respect to the matrix of items from which it is drawn. The saturation of a test is defined as the proportion of the total test variance due to inter-item covariances. A cycling process involving elimination and addition of items is followed to assure an adequate degree of independence to the keys.

3. The *Biographical Inventory BE601B* yielded the following nine keys: Mechanical, Radio, Clerical, Security-Routine, Response Set, Supervisory, Handyman, Extroversion (?), and Education.

4. Saturations of these keys ranged from .42 to .78 for the standardization sample of 1000 airmen, and from .37 to .76 for the cross-validation sample of 991 airmen. The mean of the absolute values of the intercorrelations of the keys was .15 for the standardization sample and .14 for the cross-validation sample.

5. The multiple correlations for predicting seven Air Force empirical or criterion keys from the nine homogeneous keys range from .63 to .85 for the standardization sample. Using the same regression weights in an equation for predicting empirical from homogeneous keys, the cross-validation sample yielded values from .57 to .83 for the correlation of predicted empirical key and actual value.

BIBLIOGRAPHY

1. BROGDEN, H.E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 1946, 11, Pp. 197–214.

2. CRONBACH, L.J. Test validity as a function of item difficulty. Paper delivered at 1951 meetings of APA, 1951.

3. CRONBACH, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, Pp. 297–334.

4. FERGUSON, G.A. On the theory of test discrimination. *Psychometriko*, 1949, 14, Pp. 61–68.

5. GULLIKSEN, H. The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, 1945, 10, Pp. 79–91.

6. GULLIKSEN, H. *Theory of mental tests.* New York: Wiley, 1950.

7. GUTTMAN, L. A basis for scaling qualitative data. *Amer. Sociol. Rev.*, 1944, 9. Pp. 139–150.

8. JACKSON, R.W.B., and FERGUSON, G.A. Studies on the reliability of tests. Bull. No. 12, Dept. Educ. Res., Univ. Toronto, 1941.

9. KRUDER, G.F., and RICHARDSON, M.W. The theory of the estimation of test reliability. *Psychometrika*, 1937, 2, Pp. 151–160.

10. LOEVINGER, JANE. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, 61 (4).

11. SPEARMAN, C. *The abilities of man.* New York: MacMillan, 1927.

12. TUCKER, L.R. Maximum validity of a test with equivalent items. *Psychometrika*, 1946, 11, Pp. 1–13.

Manuscript received 24 March 1952

82-1430/Scott AFB, Ill./11 July 52/900