

Air Research and Development Command

**Human
Resources
Research
Center**

*A Comparison Between the Empirical and Rational Approaches
For Keying a Heterogeneous Test*

by

Marvin H. Berkeley

**RESEARCH
BULLETIN
53-24**

A COMPARISON BETWEEN THE EMPIRICAL AND RATIONAL APPROACHES
FOR KEYING A HETEROGENEOUS TEST

Project No. 503-001-0011

By

MARVIN H. BERKELEY

6560th Research and Development Group
(Personnel Research Laboratory)
Human Resources Research Center
Air Research and Development Command
Lackland Air Force Base
Texas

RESEARCH BULLETIN 53-24
July 1953

SUBMITTED BY:

LLOYD G. HUMPHREYS
Director of Research
6560th Research and Development Group

ACKNOWLEDGMENTS

Dr. Philip H. DuBois gave general supervision and together with Dr. Jane Loevinger helped to develop the research design. Dr. Vernon W. Lemmon and Dr. Elise B. Webb acted as advisors. The author is also indebted to the many administrators, clerical assistants, and IBM technicians, who in varying degrees were involved in the extensive computational labor.

This report is a modified version of a dissertation presented to the Graduate Board of Washington University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, January 1954. The homogeneous keys were developed at Washington University under Contract AF 33 (038)-10538.

TABLE OF CONTENTS

	Page
List of Tables	iv
Introduction	1
Purpose	1
Historical Background	1
Hypotheses Tested	4
Population and Criteria	5
Preliminary Procedures	6
Restatement of the Problem	7
Homogeneous Keying	7
Empirical Keying	16
Validation of Keys	16
Validation of the Homogeneous Keys	16
Cross-Validation of the Homogeneous and Empirical Keys	21
Psychological Comparison of the Keys	27
Interpretation of Results.	32
Evaluation of the Cross-Validation	32
Empirical Versus Homogeneous Keying in a Program of Research	33
Summary and Conclusions.	34
Bibliography	36

LIST OF TABLES

Table		Page
1	First-Cycle Homogeneous Category Data	9
2	Intercorrelations of First-Cycle Homogeneous Categories . . .	10
3	Second-Cycle Homogeneous Category Data.	11
4	Intercorrelations of Second-Cycle Homogeneous Categories . .	12
5	Final Homogeneous Category Data--Third Cycle.	13
6	Intercorrelations of Independent Homogeneous Categories --Third Cycle.	14
7	Intercorrelations of Independent Homogeneous Categories by Cycles	15
8	Correlations and Summary Data of Tentative Empirical Keys and Criteria	17
9	Correlations and Summary Data of Empirical Keys and Criteria-- First Cycle.	18
10	Correlations and Summary Data of Final Empirical Keys and Criteria	19
11	Comparison Between Tentative and Final Empirical Keys	20
12	Intercorrelations of Homogeneous Keys	22
13	Multiple Correlational Data for Prediction of Final Grade in Officer Candidate School by Homogeneous Keys	23
14	Multiple Correlational Data for Prediction of Military Grade in Officer Candidate School by Homogeneous Keys.	24
15	Multiple Correlational Data for Prediction of Academic Grade in Officer Candidate School by Homogeneous Keys.	25
16	Multiple Correlational Data for Prediction of Pass/Fail in Officer Candidate School by Homogeneous Keys	26
17	Cross-Validation of Empirical and Homogeneous Keys	28
18	Comparison of the Shrinkages of the Empirical and Homogeneous Keys after Cross-Validation.	29

A COMPARISON BETWEEN THE EMPIRICAL AND RATIONAL APPROACHES FOR KEYING A HETEROGENEOUS TEST

INTRODUCTION

Purpose

The purpose of this study is to compare two approaches of keying a biographical inventory. One approach consists of the empirical derivation of keys on external criteria. This is the well-known technique of selecting from a pool of items those which yield a maximum correlation with a criterion. The other approach consists of the development of homogeneous and relatively independent keys. These keys will show high internal consistency, and the item selection will not depend upon a relationship with an external criterion.

After the homogeneous keys are validated on the same external criteria with which the empirical keys were developed, both keys are to be evaluated by means of cross-validation on a new sample. The biographical inventory to be keyed is given experimentally at present to officer candidates in the Air Force and is known by the code number, CE608C. This study has implications for the construction, analysis, and use of such tests as they apply in educational, vocational, and personality guidance.

Historical Background

Empirical Keying

Twenty-three empirical methods have been described by Long and Sandiford (18). Gulliksen (11, p. 364) notes that of these, nearly all overlook the theoretical aspects of the reliability and validity of the total test with a few notable exceptions (1, 13, 20, 22, 24).

Apart from those methods which were listed in the Long and Sandiford survey, others have since been proposed. Several of these are presented as representative procedures of empirical keying.

Horst (12) has devised a method which involves the computation of the mean criterion score and the mean total test score of all subjects who answered correctly on any particular item. Through plotting, this method retains the items with the largest index. The author claims this method is less time-consuming and, at the same time, yields at least as high validities as his method of successive residuals.

Flanagan (7) suggests a method of item selection in which a nucleus of the most valid items are first selected, and items are added to or subtracted from this nucleus by comparing the item-nucleus correlation of each

item with the item-criterion correlation. The items having a higher correlation with the criterion than with the nucleus are retained, while the others are dropped. The cycle can be repeated, but Flanagan notes that only a small increment of improvement results from additional cycles.

Gleser and DuBois (8) have developed a method very similar to that of Flanagan. However, they utilize the item-criterion and item-test correlations to compute an index for each item of the form:

$$r_{jc} - \frac{r_{jt} + \frac{\sigma_j}{2\sigma_t}}{2}$$

This index provides a correction for whether or not the item is included in the nucleus "t," and also takes into account the changes in item-total correlations which result after the first selection is made. It provides an exact criterion of how many items to retain in the final test.

Dailey (14) has presented a relatively recent method for keying biographical data empirically. This method grew out of the inadequacy of the method of selecting those responses with validity coefficients above a given level of significance. In this method, called the "pattern of response method," all possible responses are correlated with a criterion, yielding continua of correlations with multiple-choice items. Those items for which the correlations show a consistent direction are keyed. Positive or negative unit weights are assigned according to the sign of the coefficient, and only the extremes are usually keyed. When cross-validated with subsequent samples, this method resulted in less shrinkages and greater validity than the method of simply choosing significant items.

Each one of the preceding techniques represents the empirical approach to keying items. With slight modifications or combinations of two or more principles inherent in each method, any one may qualify as the representative of the empirical approach for the purposes of this study. Systematic comparisons of the methods of item analysis (8, 11, 15, 18, 19, 23) have not given much satisfaction for selection of the best method. The choice of method seems to depend upon the labor which is involved for the obtained increase in validity, stability of the validity coefficient for subsequent samples, and the ultimate purpose with which the test will be used. It is to be noted that with the exception of the "pattern of response method," there is little evidence for the greater stability of any one method over any other. It is to be emphasized, furthermore, that almost all the methods have some points in common with others. At least one fact, however, provides a basis for the selection of the method to be used in this study. It has long been known that given n items with identical validities, the two items having the lowest correlations with each other will predict the criterion better than will any other of the possible pairs of items. In other

words, the items which are selected for inclusion in an empirical key should lend unique valid variance as far as possible. Therefore, in the development of an empirical key, the intercorrelations between the items of the key should be considered. This may be done directly by considering the item relationships, or indirectly by considering the item-total test relationships. The Gleser and DuBois method of maximizing test validity (6) was selected on the basis of the latter consideration.

Homogeneous Keying

Zubin (26) was perhaps the first who applied different methods for computing item-total relationships in an attempt to develop a homogeneous test. He noted that with the lack of suitable external criteria, as is often the case with personality inventories, proceeding by means of the internal consistency of the test is the next best approach.

Factor analytic techniques have been combined with item analysis on such tests as the Guilford-Martin "Inventory of Factors GAMIN," and the Guilford "Inventory of Factors STDCR." The major criticisms directed against these tests are their lack of validation data and their laborious statistical computation. At the same time, substantial overlap of the scales was developed; in some cases the scales were intercorrelated as high as the .70's. Favorable criticism of the technique centers around the general advances given to test construction, as well as their independence of obsolete and unreliable psychiatric classification (4, pp. 80, 82). This latter criticism may, of course, be given for any of the methods which aim toward the development of homogeneous tests.

Loevinger (16, 17) conceives of homogeneity essentially as the average correlation of items within the test. She presents two coefficients designed to give the degree of homogeneity between any two items and the homogeneity of the test respectively. Cronbach and Damrin (5), however, have criticized the use of Loevinger's coefficients as being markedly dependent on the difficulties of items, and, furthermore, they demonstrated that the coefficients do not apply when the relationships between items are low. It should be noted that the concept of homogeneity is dependent on the type of test involved. In ability tests item relationships are high, whereas in personality-type tests intercorrelations between items are characteristically moderate or low. Cronbach and Damrin showed lastly that the Kuder-Richardson Formula 20, or its derivative "phi bar," was sufficient to show the equivalence of the items up to the point where the correlations between items of equal difficulty rise to .80 and .90. This formula, which is the mean of all possible split-half coefficients of the test, might be directly interpreted as the proportion of the test variance that is contributed by the common factors among the items. This systematic use of Kuder-Richardson Formula 20 represents an untested though fairly laborious approach for constructing a homogeneous test.

Another method which results in a homogeneous test is referred to as "maximizing test saturation" by DuBois, Loevinger, and Gleser (6). Briefly described, this method takes into account the ratio of common factor variance that the items contribute to the total variance of a test. This ratio has been titled "the saturation of the test." When items are added successively to a nucleus of three or four highly intercorrelated items, so as to maximize the saturation, this should result in a homogeneous test. Moreover, if one were to start with nuclei that have little in common, the keys that are subsequently developed should be relatively independent. This method was selected to represent the homogeneity approach.

Examination of the history of keying tests homogeneously reveals little application to test validation. What has been done has been carried out only for personality-type tests where external criteria are unsuitable or lacking. Keying empirically has been carried out generally whenever a choice between the two was to be made. It would appear that a crucial study would involve the comparison of two rigorous methods representing each approach for keying the same biographical inventory. This, in short, is the over-all purpose of this study.

HYPOTHESES TESTED

The following hypotheses are tested by this study:

1. The empirical keys will contain higher correlations with the criteria than the homogeneous keys on the developmental sample. In the first place, Biographical Inventory CE608C was developed by the inclusion of those items which were shown to be valid for prediction of CCS success. Secondly, the empirical keys included items on the basis of their specific contribution to the prediction of an external criterion. On the other hand, homogeneous keys are constructed solely on the basis of the internal consistency of the items which may or may not be related to the criterion. The empirical keys are expected, therefore, to be characteristically more valid than the homogeneous keys.

2. The empirical keys will show a greater shrinkage and a lower validity than the homogeneous keys. The items of the homogeneous keys tend to duplicate each other, resulting in the probable cancellation of chance errors. By contrast the empirical keys will approach the heterogeneity of the criteria they are designed to predict. For this reason homogeneous keys can be expected to be generally more reliable than empirical keys. Guilford (9) has also pointed out that factorially impure tests (empirical keys) contain variance that is unrelated to the criterion. This invalid variance adds spuriously to the validity when chance deviations are optimally weighted, and this serves to lower a cross-validity as would a like amount of error variance. It would, therefore, be expected that these two factors would result in a greater shrinkage and lower cross-validity for

the empirical keys than for the homogeneous keys.

3. The homogeneous keys will be psychologically meaningful and easy to interpret, and the empirical keys will be psychologically complex and difficult to interpret. Insofar as the method of homogeneous keying is successful in its primary goal of maintaining psychological purity, these keys should be simple to interpret. Since the empirical keys will be composed of a multitude of factors, it should be difficult to know which of the specific factors to invoke and to what degree in order to explain a person's score. In this regard Guilford and Lacey (10, p. 881) state: "This (empirical) procedure would seem merely to result in an extension of our ignorance to new valid territory, rather than to increase our knowledge of why tests are valid and therefore to improve our control over validity already achieved."

POPULATION AND CRITERIA

The first step in keying by either the homogeneous or empirical approach is to obtain the sample with which such keys are to be developed. The sample on which the homogeneous keys were to be developed was designated, Sample A. It was obtained by selecting every third paper out of a total pool of basic airmen who were administered CE608C during November 1950 until a total of 1000 papers was obtained. Most of the airmen were in their second week of military experience. These 1000 papers were then scanned for completeness and correct scoring. The sampling process and examination of papers was continued until a total of 1000 valid papers was obtained.

The sample on which the empirical keys were to be developed and on which the homogeneous keys were to be validated was designated, Sample B. This sample was composed of all available male graduates and eliminees of Officer Candidate School (OCS) Classes 50-A, 50-B, and 50-C, and totaled 336 graduates and 78 eliminees. The sample with which the empirical keys and the homogeneous keys were cross-validated was designated, Sample C. This sample included all available male graduates and eliminees of OCS Classes 51-A and 51-B and totaled 306 graduates and 29 eliminees.

The fact that the homogeneous keys were developed on airmen and validated on officer candidates, while the empirical keys were developed on officer candidates, may represent a serious limitation in the study. The use of airmen was necessitated by the lack of a sufficient number of officer candidates who had been administered CE608C and for whom criteria data were available. Since the basic personality variables, as elicited by CE608C, may have been somewhat different for the airmen and officer candidate populations, this may have served to reduce the validity of the homogeneous keys. A comparison of the two keys should be interpreted, therefore, with this limitation in mind.

The study is dependent upon the nature of the criteria which included the following:

1. Pass/fail. This criterion was determined by splitting the CCS classes into those who graduated and those who failed to complete the course. Failure to complete CCS was due to either low over-all grades or resignation.
2. CCS military grades. These were ratings on military potential made by the Tactical Officer in charge of each flight. Each flight was composed of 25 officer candidates.
3. CCS academic grades. This criterion was determined by differentially weighting (dependent upon the number of hours devoted to each course) objective achievement test scores in the various academic subjects taught. These subjects included personnel methods, supply, administration, military law, etc.
4. CCS final grades. These grades were obtained by equally weighting the academic and military grades into a composite.

PRELIMINARY PROCEDURES

The first step of both the homogeneous and empirical approach was the dichotomization and weighting of items. A decision had to be made for the most meaningful dichotomy of choices on a five-choice continuum. One part of the dichotomy was to be given unity weight which would arbitrarily assign a zero weight to the other part of the dichotomy.

In order to carry out the dichotomization and weighting of items, the first procedure was a random selection of a subsample of 250 papers from Sample A. An item count was then obtained of all possible answers. On the basis of the item count and logical consideration of judges as to the part the item might later play in a priori keys, all the items were split dichotomously, approximating the 50-50 split as far as possible. In a few cases where the items were "double-barrelled" or "bifurcated," thus presenting two possible splits with distinctly separate interpretations, two items were developed out of one.

It soon became apparent that many items had to be eliminated from further consideration because the items were concerned with Air Force experience. These items would obviously not yield any appreciable valid variance, since, as pointed out before, almost all the examinees were in their second week as airmen. Out of an original 327 possible items available for keying, there now remained 183. These preliminary procedures were common for both the empirical and homogeneous approach. From this point both methods proceed in divergent directions.

RESTATEMENT OF THE PROBLEM

This study was to be carried out in accordance with the following objectives:

1. Derivation of homogeneous and relatively independent keys for Biographical Inventory CE608C on Sample A. This was to be done by the method of maximizing test saturation.
2. Intercorrelation of homogeneous keys on Sample B.
3. Validation of each homogeneous key on each criterion of Sample B to include: final OCS grade, military OCS grade, academic OCS grade, and pass/fail in OCS.
4. Computation of beta coefficients for each homogeneous key for each criterion and computation of the coefficients of multiple correlation with each criterion.
5. Scoring of Sample C on the homogeneous keys and weighting of each homogeneous key score by beta weights established in each multiple regression formula, as determined from Sample B.
6. Summation of weighted homogeneous key scores to yield a predicted criterion score.
7. Correlations of predicted criterion scores with actual criterion scores for Sample C.
8. Derivation of the empirical keys for each criterion of Sample B by the Gleser-DuBois method for maximizing test validity.
9. Scoring of Sample C on each empirical key, and correlations of empirical key scores against criterion scores.
10. Comparison of the validities and cross-validities of both sets of keys, and evaluation of the relative difficulties and characteristics of each method.
11. Psychological comparison of both sets of keys.

HOMOGENEOUS KEYING¹

The first step of homogeneous keying was an a priori categorization of the 183 items available for keying by three judges. This resulted in

¹ A detailed theoretical and methodological presentation of the method for maximizing test saturation may be found in DuBois, Levinger, and Gleser (6).

the formation of 13 categories which showed promise of common factor content. Of these 13 categories, four were combined since it was felt that each of the four might possess a fairly high relationship with its respective paired member.

Having carried out the procedure for maximizing test saturation, 13 first-cycle categories were derived from the a priori categories. The first-cycle categories included a total of 129 items of which three items were included in two categories. From the residual number of 57 unplaced items, one additional category of 11 items was developed. Each category was named following inspection of the item content. The category data, including the name, mean, variance, and saturation are given in Table 1; and the category intercorrelations are given in Table 2. Each homogeneous category is identified by a letter which indicates the a priori cluster. Where more than one category was derived from the a priori cluster, a subscript accompanies the letter, indicating the order of category evolution.

As may be noted in Table 2, one of the categories, Aggressiveness, seemed to resemble a general factor, since it correlated high (above .34) with one-half of the other categories. Since it was intended to develop independent categories with as many items as possible, it was decided to put the eight items comprising this category back into the general pool of unused items and to reconsider them following the development of independent categories. Two of the eight items were included in the independent categories in a later cycle.

Category intercorrelations in Table 2 were now examined to determine the feasibility of combining two or more highly correlated categories into a single matrix. When the general factor category was removed from consideration, seven correlations remained which ranged from .35 to .49. It was decided to delay any combinations of categories until an inspection of the intercorrelations of the completed first-cycle categories, at which time all items would have been correlated with all categories. Since the number of correlations exceeding .35 dropped from seven to two as a result of the removal of eight items from first-cycle categories, it was decided to continue the cycling without combining any first-cycle categories.

In order to achieve greater independence of categories without much loss of saturation, categories were revised in a second and a third cycle. Tables 3 and 4 and Tables 5 and 6 present the category data and category intercorrelations with Cycles 2 and 3, respectively. Table 6 also includes the data for the revised general factor category following the inclusion of seven items which added to the saturation. Table 7 presents a comparison of the independence of categories as a result of the cycling process. It may be noted that in order to achieve a decrease of average correlation between categories of .05, 15 per cent of the total possible number of items had to be dropped from Cycle 1 to Cycle 3.

Table 1
First-Cycle Homogeneous Category Data
(Sample: 1000 basic airmen)

Category	No. of Items	Mean	Vari- ance	Satur- ation ^c
A Mechanical Aptitude	9	6.12	3.16	.48
B ₁ Athletic Experience	13	6.25	9.49	.70
B ₂ Childhood Games	8	6.21	3.30	.60
C Playboy ^a	12	4.22	3.98	.49
D Socio-Economic	17	9.04	14.08	.74
E Schizoid ^a	6	1.96	1.94	.36
F ₁ Parental Criticism	11	6.51	8.02	.68
F ₂ Extroversion	13	5.33	8.13	.67
F ₃ Aggressiveness	8	4.74	3.34	.50
G Itinerant ^a	6	2.63	2.54	.50
H Scholarship	13	4.32	6.94	.62
I Societal Acceptance	13	7.50	6.71	.59
J Childhood Responsibility ^b	11	5.13	5.82	.57

^a These category names were changed from a priori names following examination of item content.

^b This category was developed out of the residual items ($n = 57$) unplaced in Cycle 1.

^c Saturation = $2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n C_{ij} \div \sum_{i=1}^n V_i + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n C_{ij}$ where C_{ij} = covariance between any two items and V_i = variance of any item i .

Table 2

Intercorrelations of First-Cycle Homogeneous Categories

(N = 1000)

Category ^a	Mean	SD	Satur- ation	A	B ₁	B ₂	C	D	E	F ₁	F ₂	F ₃	G	H	I	J
A Mech Aptitude	6.12	1.78	.48		.19	.33	.00	.24	-.01	.01	.08	.20	.19	-.05	.10	.35
B ₁ Athletic Exp	6.25	3.08	.70	.19		.30	.15	.34	-.13	-.08	.39	.60	.19	.14	.13	.30
B ₂ Childhd Games	6.21	1.82	.60	.33	.30		.26	.39	-.14	.03	.21	.42	.19	.16	.16	.40
C Playboy	4.22	2.00	.49	.00	.15	.26		.15	-.31	-.15	.02	.38	.23	-.07	-.03	.17
D Socio-Teen	5.04	3.75	.74	.24	.34	.39	.15		-.08	.05	.44	.35	.23	.31	.25	.41
E Schizoid	1.96	1.39	.36	-.01	-.13	-.14	-.31	-.00		.17	-.07	-.24	-.07	.02	.05	-.07
F ₁ Parent Crit	6.51	2.38	.67	.01	-.08	.03	-.15	.05	.17		.04	-.11	-.02	.15	.15	.02
F ₂ Extroversion	5.33	2.83	.67	.08	.39	.21	.02	.44	-.07	.04		.40	.16	.49	.30	.51
F ₃ Aggressive	4.74	1.83	.50	.20	.60	.42	.38	.35	-.24	-.11	.40		.21	.14	.09	.34
G Itinerant	2.63	1.59	.50	.19	.19	.19	.23	.23	-.07	-.02	.16	.21		.07	.11	.25
H Scholarship	4.32	2.64	.62	-.05	.14	.16	-.07	.31	.02	.15	.49	.14	.07		.34	.22
I Societal Accep	7.50	2.59	.59	.10	.13	.16	-.03	.25	.05	.15	.30	.09	.11	.34		.17
J Childhd Respon	5.13	2.41	.57	.35	.30	.45	.17	.41	-.07	.02	.31	.34	.26	.22	.17	

^a See Table 1 for full category title.

Table 3
Second-Cycle Homogeneous Category Data
(N = 1000)

<u>Category</u>	<u>No. of items</u>	<u>Mean</u>	<u>Vari- ance</u>	<u>Satur- ation</u>
A Mechanical Aptitude	7	4.89	2.07	.44
B ₁ Athletic Experience	10	4.77	6.50	.67
B ₂ Childhood Games	7	5.52	2.59	.57
C Playboy	11	3.53	3.67	.45
D Socio-Economic	14	7.55	10.60	.71
E Schizoid	6	1.96	1.94	.36
F ₁ Parental Criticism	11	6.51	8.02	.68
F ₂ Extroversion	13	5.33	8.13	.67
G Itinerant	6	2.63	2.54	.50
H Scholarship	11	3.81	5.66	.60
I Societal Acceptance	14	7.87	7.48	.60
J Childhood Responsibility	7	3.32	2.99	.46

Table 4

Intercorrelations of Second-Cycle Homogeneous Categories

(N = 1000)

Category ^a	Mean	SD	Saturation	A	B ₁	B ₂	C	D	E	F ₁	F ₂	G	H	I	J
A Mech Aptitude	4.89	1.44	.44		.15	.19	-.07	.16	.00	-.02	.01	.16	.13	-.11	.25
B ₁ Athletic Exp	4.77	2.55	.67	.15		.15	.15	.26	-.13	-.06	.33	.16	.09	.11	.21
B ₂ Childhd Games	5.52	1.61	.57	.19	.15		.17	.31	-.11	.04	.18	.17	.12	.19	.39
C Playboy	3.53	1.91	.45	-.07	.15	.17		.17	-.24	-.16	.05	.24	-.05	-.03	.17
D Socio-Econ	7.55	3.26	.71	.16	.24	.31	.17		-.08	.05	.37	.20	.25	.22	.51
E Schlzoid	1.96	1.39	.36	.00	-.13	-.11	-.24	-.08		.17	-.06	-.07	.02	.07	-.07
F ₁ Parent Crit	6.51	2.83	.63	-.02	-.04	.04	-.16	.09	.17		.04	-.02	.16	.15	.01
F ₂ Extroversion	5.33	2.83	.67	.01	.33	.18	.05	.37	-.06	.04		.16	.46	.30	.23
G Itinerant	2.63	1.59	.50	.16	.13	.17	.24	.20	-.07	-.02	.16		.06	.11	.12
H Scholarship	3.81	2.38	.60	.13	.09	.12	.05	.23	.02	.16	.46	.06		.33	.14
I Societal Accep	7.87	2.74	.60	-.11	.11	.15	-.03	.22	.06	.15	.30	.11	.33		.14
J Childhd Respon	3.32	1.73	.46	.25	.21	.30	.13	.31	-.07	.01	.23	.22	.14	.14	

^a See Table 3 for full category title.

Table 5
Final Homogeneous Category Data--Third Cycle
(N = 1000)

<u>Category</u>	<u>No. of Items</u>	<u>Mean</u>	<u>Vari- ance</u>	<u>Satur- ation</u>
A Mechanical Aptitude	7	4.89	2.07	.44
B ₁ Athletic Experience	10	4.77	6.50	.67
B ₂ Childhood Games	7	5.52	2.59	.57
C Playboy	13	3.96	4.53	.52
D Socio-Economic	13	7.23	9.52	.71
E Schizoid	6	1.96	1.94	.30
F ₁ Parental Criticism	11	6.51	8.02	.68
F ₂ Extroversion	11	4.64	6.40	.65
G Itinerant	6	2.63	2.54	.50
H Scholarship	8	2.75	3.55	.54
I Societal Acceptance	13	7.51	6.66	.58
J Childhood Responsibility	6	2.60	2.44	.44
F ₃ Aggressiveness	15	8.59	8.96	.64

Table 6

Intercorrelations of Independent Homogeneous Categories--Third Cycle

(N = 1000)

Category ^a	Mean	SD	Saturation	A	B ₁	B ₂	C	D	E	F ₁	F ₂	G	H	I	J
A Mech. Aptitude	4.89	1.44	.44		.15	.20	-.11	.15	.00	-.02	.01	.15	-.12	.07	.24
B ₁ Athletic Exp	4.77	2.55	.67	.15		.16	.10	.26	-.13	-.06	.32	.18	.11	.10	.19
B ₂ Childhd Games	5.52	1.61	.57	.20	.16		.15	.32	-.12	.05	.18	.18	.10	.14	.31
C Playboy	3.96	2.13	.52	-.11	.10	.15		.16	-.25	-.16	.04	.20	-.07	-.07	.13
D Socio-Econ	7.23	3.09	.71	.15	.26	.32	.16		-.03	.05	.33	.21	.18	.22	.29
E Schizoid	1.96	1.39	.36	.00	-.13	-.12	-.25	-.03		.17	-.06	-.07	.00	.06	-.06
F ₁ Parent Crit	6.51	2.83	.68	-.02	-.06	.05	-.16	.05	.17		.03	-.02	.15	.15	.01
F ₂ Extroversion	4.64	2.53	.65	.01	.32	.18	.04	.33	-.06	.03		.15	.38	.27	.22
G Itinerant	2.63	1.59	.50	.13	.18	.18	.20	.21	-.07	-.02	.15		.06	.12	.23
H Scholarship	2.75	1.88	.54	-.12	.11	.10	-.07	.18	.00	.15	.38	.06		.29	.11
I Societal Accep	7.51	2.58	.58	.07	.10	.14	-.07	.22	.06	.15	.27	.12	.29		.14
J Childhd Respon	2.60	1.56	.42	.24	.19	.31	.13	.29	-.06	.01	.22	.23	.11	.14	

^a See Table 5 for full category title.

Table 7
Intercorrelations of Independent Homogeneous
Categories By Cycles

<u>r^a</u>	<u>Cycle 1</u>	<u>Cycle 1A^b</u>	<u>Cycle 2</u>	<u>Cycle 3</u>
.60 - .64	1			
.55 - .59				
.50 - .54				
.45 - .49	2	2	1	
.40 - .44	4	2		
.35 - .39	5	3	1	1
.30 - .34	10	9	6	4
.25 - .29	3	3	2	5
.20 - .24	8	5	8	7
.15 - .19	15	15	17	15
.10 - .14	3	6	10	14
.05 - .09	12	11	12	12
.00 - .04	<u>10</u>	<u>10</u>	<u>9</u>	<u>8</u>
Total correlations	78	66	66	66
Total items used	140	132	117	111
Number of items added and/or dropped	___	Dropped 8	Added 2 Dropped 17	Added 2 Dropped 8
Per cent of pos- sible items used	73	68	61	58
Average correla- tion	.201	.183	.153	.146

^a Signs of intercorrelations are omitted.

^b Category "Aggressiveness" which appeared as a general factor in Cycle 1 was dropped for Cycle 1A.

EMPIRICAL KEYING²

The tentative empirical keys were to be formed by including those items whose correlation with a criterion was significant at the .01 level of confidence. Examination of the correlations revealed that 16, 18, and 20 items qualified for inclusion into the final grade, military grade, and academic grade keys, respectively. However, it was also noted that only seven items qualified at the .01 level of significance for inclusion into the pass/fail key. It was decided to lower the requirement for including an item in the pass/fail key to the .05 level of significance. This decision resulted in the addition of five more items or a total of 12 items in the pass/fail key.

Sample B answer sheets were scored on the tentative empirical keys, and these scores were correlated against their respective criterion. These correlations and other summary data of the tentative empirical keys, including the items in the keys, means, and standard deviations are presented in Table 8.

First-cycle empirical keys were now developed by the Gleser and DuBois method of maximizing test validity. The answer sheets were scored on the first-cycle keys, and these scores were correlated with their respective criterion. The key-criterion correlations and summary data for first-cycle keys are presented in Table 9.

Since the magnitude of each first-cycle key-criterion correlation increased by at least four correlation points, a second cycle was carried out. It was noted at the completion of the second cycle that the key-criterion correlations increased only slightly above those of the preceding cycle, and, therefore, no additional category refinement seemed necessary. The key-criterion correlations and summary data for second-cycle categories are presented in Table 10. The comparative changes from the tentative empirical keys to the final keys are presented in Table 11. It was now possible to score a new sample on both the empirical and homogeneous keys and compare their respective validities.

VALIDATION OF KEYS

Validation of the Homogeneous Keys

It has been pointed out that the homogeneous keys were developed independently of any external criteria. Prior to a cross-validation, therefore, it was necessary to obtain the intercorrelation of the keys and the validities and beta weights for those criteria which were used to develop the empirical keys. For control purposes, the same validating sample with which the empirical keys were developed was utilized to obtain these data. This

² For a detailed theoretical and methodological presentation of the method of maximizing test validity, cf. Gleser and DuBois (8).

Table 8
Correlations and Summary Data of Tentative
Empirical Keys and Criteria

<u>Criteria or key</u>	<u>N</u>	<u>Items in key</u>	<u>Mean</u>	<u>SD</u>	<u>r</u>
Final grade ^a	336		4.99	1.97	
Final grade key	336	16	8.62	2.04	.39
Military grade ^a	336		5.08	1.89	
Military grade key	336	18	11.18	2.56	.50
Academic grade ^a	336		5.03	1.96	
Academic grade key	336	20	11.57	2.63	.50
Pass/fail	414		.81	.39	
Pass/fail key	414	12	6.77	1.76	.45 ^b

^a Standardized in stanine units.

^b Biserial correlation coefficient, where $p = .81$ and $q = .19$.

Table 9

Correlations and Summary Data of Empirical
Keys and Criteria--First Cycle

<u>Criteria or key</u>	<u>N</u>	<u>Items in key</u>	<u>Mean</u>	<u>SD</u>	<u>r</u>
Final grade ^a	336		4.99	1.97	.43
Final grade key	336	29	17.22	3.03	
Military grade ^a	336		5.08	1.89	.54
Military grade key	336	32	21.13	3.34	
Academic grade ^a	336		5.03	1.96	.56
Academic grade key	336	34	20.16	3.38	
Pass/fail	414		.81	.39	.50 ^b
Pass/fail key	414	17	11.78	2.17	

^a Standardized in stanine units.

^b Biserial correlation coefficient, where $p = .81$ and $q = .19$.

Table 10
Correlations and Summary Data of Final
Empirical Keys and Criteria

<u>Criteria or key</u>	<u>N</u>	<u>Items in key</u>	<u>Mean</u>	<u>SD</u>	<u>r</u>
Final grade ^a	336		4.99	1.97	
Final grade key	336	39	21.10	3.51	.43
Military grade ^a	336		5.08	1.89	
Military grade key	336	40	27.86	3.67	.54
Academic grade ^a	336		5.03	1.96	
Academic grade key	336	39	23.47	3.54	.58
Pass/fail	414		.81	.39	
Pass/fail key	414	19	12.49	2.50	.51 ^b

^a Standardized in stanine units.

^b Biserial correlation coefficients, where $p = .81$ and $q = .19$.

Table 11

Comparison Between Tentative
And Final Empirical Keys

<u>Key</u>	<u>Cycle</u>	<u>No. of items</u>	<u>Per cent of possible items used</u>	<u>Increase in r with criterion</u>
Final grade	Cycle 1	16	09	.04
	Final	39	20	
Military grade	Cycle 1	18	09	.04
	Final	40	21	
Academic grade	Cycle 1	20	10	.08
	Final	39	20	
Pass/fail	Cycle 1	12	06	.06
	Final	19	10	

sample, known as Sample B, included 35+ graduates and 78 eliminees of CCS Classes 50-A, 50-B, and 50-C.

Table 12 presents the intercorrelations of the homogeneous keys based upon the independent Sample B. The magnitudes of the intercorrelations were strikingly similar to those obtained on Sample A, and the average correlation of the matrix, minus the general factor category, increased only eight points in the third decimal. Also noteworthy was the fact that the general factor now cut across the categories less than it did in Sample A. This fact can probably be attributed to the addition of seven items to the general factor, since the category was not re-correlated with the other categories following its final revision in the third cycle. A last point to be noted in Table 12 was the shrinkage of some saturation coefficients. (Compare these saturations with those in Table 7.) Shrinkage of the saturation coefficient occurs for the same reason as for a correlation coefficient: the error factor in the first sample is weighted in favor of the original keying, and since error variance does not reproduce itself in subsequent administrations, additional error appears, and the saturation or correlation coefficient diminishes. It should be noted that a shrunken saturation coefficient represents a truer estimate of homogeneity.

Having obtained a truer estimate of the intercorrelations of the homogeneous keys and their separate validities on four criteria, four sets of beta weights and four multiple correlations were computed. The detailed data for these multiple correlations, including the homogeneous keys comprising the predictor composite, beta weights, validities, and multiple R's, are given in Tables 15 through 16.

Cross-Validation of the Homogeneous and Empirical Keys

Sample C, which was composed of 306 graduates and 31 failures of CCS Classes 51-A and 51-B, was scored on each of the four empirical keys and on the 13 homogeneous keys. Three Pearson product-moment correlations were obtained for the empirical keys against their respective CCS grades, and one biserial correlation coefficient was computed for pass/fail on its empirical key. These correlations represented the cross-validities of the empirical keys. In order to obtain the multiple validities of the homogeneous keys, the raw scores of the keys comprising the predictor composite were weighted by their particular regression weight. These weighted scores were summed along with the constant term to give a composite predicted criterion score for each subject. Each predicted criterion score then was correlated against the subject's obtained criterion score to give the multiple-validity correlation. A comparison of the data comprising the cross-validation is given in Table 17.

One of the most significant comparisons of the two keys to be made in this study was between validities of the empirical keys and the multiple validities of the homogeneous keys. Table 17 gives the critical ratios for the differences. Inspection of Table 17 reveals that in cross-validation

Table 12

Intercorrelations^a of Homogeneous Keys

(Sample: 412 officer candidates of Classes 50-A, 50-B, and 50-C)

Category ^b	Mean	SD	Satur- ation	A	B ₁	B ₂	C	D	E	F ₁	F ₂	G	H	I	J	K
A Mech Aptitude	4.71	1.52	.47		.24	.18	-.03	.21	-.01	-.06	.08	.08	.01	.06	.28	.24
B ₁ Athletic Exp	5.42	2.51	.67	.24		.10	-.11	.09	-.01	-.08	.22	.11	.11	.07	.12	.61
B ₂ Childhd Games	6.23	1.06	.40	.18	.10		.21	.18	-.06	-.08	.04	.11	-.03	.03	.15	.36
C Playboy	4.54	2.13	.52	-.03	-.11	.21		.11	-.18	-.21	-.19	.09	-.13	-.16	.00	.31
D Socio-Econ	8.59	2.69	.67	.21	.09	.18	.11		.07	.01	.23	.07	.09	.15	.34	.19
E Schizoid	1.90	1.33	.31	-.01	-.01	-.06	-.18	.07		.21	-.03	-.02	.06	.02	-.04	-.18
F ₁ Parent Crit	7.27	2.62	.66	-.06	-.08	-.08	-.21	.01	.21		.02	-.09	.14	.16	-.04	-.21
F ₂ Extroversion	6.75	2.02	.54	.08	.22	.04	-.19	.23	-.03	.02		.17	.15	.12	.19	.20
G Itinerant	3.87	1.44	.41	.08	.11	.11	.09	.07	-.02	-.09	.17		.03	.04	.20	.23
H Scholarship	4.12	1.76	.43	.01	.11	-.03	-.13	.09	.06	.14	.15	.03		.14	.05	.01
I Societal Accep	9.34	2.08	.47	.06	.07	.03	-.16	.15	.02	.16	.12	.04	.14		.12	-.03
J Childhd Respon	3.20	1.43	.37	.28	.12	.15	.00	.34	-.04	-.04	.19	.20	.05	.12		.25
K Aggressiveness	10.61	2.41	.54	.24	.61	.36	.31	.19	-.18	-.21	.20	.23	.01	-.03	.25	

^a Mean Intercorrelation (excluding the general factor, Aggressiveness) = .15.^b See Table 5 for full category title.

Table 13

Multiple Correlational Data For Prediction of
Final Grade of Officer Candidate School
By Homogeneous Keys

(Sample: 414 officer candidates of Classes 50-A, 50-B, and 50-C)

<u>Homogeneous key</u>	<u>Validity</u>	<u>Beta weights</u>							
	<u>r</u>	<u>8</u>	<u>7</u>	<u>6</u>	<u>5</u>	<u>4</u>	<u>3</u>	<u>2</u>	<u>1</u>
1. Athletic Experience	.14	.27	.16	.15	.16	.14	.14	.15	.14
2. Parental Criticism	.10	.12	.14	.13	.14	.11	.11	.11	
3. Childhood Responsibility	.08	.11	.10	.06	.06	.06	.07		
4. Itinerant	.07	.06	.04	.04	.04	.05			
5. Playboy	.07	.19	.15	.11	.11				
6. Scholarship	.05	.04	.04	.03					
7. Socio-Economic	-.05	-.11	.12						
8. Aggressiveness	.04	-.18							
Multiple R		.28	.25	.23	.22	.20	.19	.18	.14

Table 14

Multiple Correlational Data For Prediction of
Military Grade in Officer Candidate School
By Homogeneous Keys

(Sample: 414 officer candidates of Classes 50-A, 50-B, and 50-C)

<u>Homogeneous key</u>	<u>Validity</u>	<u>Beta weights</u>							
	<u>r</u>	<u>8</u>	<u>7</u>	<u>6</u>	<u>5</u>	<u>4</u>	<u>3</u>	<u>2</u>	<u>1</u>
1. Scholarship	.25	.26	.26	.25	.25	.25	.25	.23	.25
2. Parental Criticism	.13	.13	.13	.14	.14	.14	.14	.10	
3. Playboy	.13	.15	.15	.16	.17	.17	.19		
4. Childhood Games	.11	.11	.11	.11	.09	.10			
5. Itinerant	.07	.07	.06	.06	.05				
6. Mechanical Aptitude	-.06	-.05	-.06	-.07					
7. Athletic Experience	-.04	-.05	-.05						
8. Extroversion	-.04	-.05							
Multiple <u>R</u>		.35	.35	.35	.34	.33	.32	.27	.25

Table 15

Multiple Correlational Data For Prediction of
Academic Grade in Officer Candidate School
By Homogeneous Keys

(Sample: 414 officer candidates of Classes 50-A, 50-B, and 50-C)

Homogeneous Key	Validity									
	r	9	8	7	6	5	4	3	2	1
1. Scholarship	.19	.17	.17	.17	.17	.18	.19	.19	.17	.19
2. Parental Criticism	.13	.14	.14	.15	.15	.16	.15	.14	.11	
3. Playboy	.12	.22	.21	.15	.17	.18	.16	.17		
4. Itinerant	.09	.05	.10	.08	.08	.08	.09			
5. Athletic Experience	.08	.20	.19	.10	.10	.08				
6. Mechanical Aptitude	-.07	-.10	-.08	-.09	-.09					
7. Childhood Games	.07	.05	.05	.05						
8. Aggressiveness	.03	-.17	-.15							
9. Childhood Responsibility	.03	.05								
Multiple R		.33	.32	.31	.31	.30	.29	.27	.22	.19

Table 16

Multiple Correlational Data For Prediction of
Pass/Fail in Officer Candidate School
By Homogeneous Keys

(Sample: All officer candidates of Classes 50-A, 50-B, and 50-C)

Validity		Beta weights											
Homogeneous key	r	12	11	10	9	8	7	6	5	4	3	2	1
1. Societal Accep	.17	.19	.18	.13	.12	.13	.18	.15	.18	.18	.19	.17	.17
2. Childhd Games	.09	.03	.03	.03	.03	.05	.04	.06	.03	.05	.06	.03	
3. PlayJoy	.09	.07	.07	.03	.08	.12	.12	.13	.12	.10	.11		
4. Socio-Econ	.10	.04	.04	.03	.03	.03	.03	.03	.04	.05			
5. Schizoid	.08	.12	.11	.12	.12	.11	.11	.10	.10				
6. Scholarship	.09	.08	.07	.03	.08	.07	.07	.07					
7. Aggressiveness	.08	.19	.17	.16	.16	.05	.05						
8. Mech Aptitude	.01	-.01	-.01	-.01	-.01	.02							
9. Athletic Exp	-.03	-.14	-.14	-.14	-.14								
10. Parent Crit	.04	.01	.01	.01									
11. Extroversion	-.02	-.04	-.06										
12. Itinerant	-.04	-.06											
Multiple R	.29	.29	.29	.28	.28	.26	.26	.26	.25	.22	.22	.19	.17

there was a tendency for the homogeneous keys to predict two of the criteria better than the empirical keys, and the empirical keys to predict the other two criteria better than the homogeneous keys. Since none of these differences was significant, it is concluded that, insofar as these data are concerned, neither the empirical nor the homogeneous method of keying proved superior.

A second important comparison to be made between the two keys was the comparison of the shrinkages. Table 18 presents the relevant data. It may be noted that for all four criteria the shrinkages of the empirical keys were significantly greater than zero beyond the .01 level of confidence. The shrinkage resulting from the cross-validation of the homogeneous keys on only one criterion, academic grade, was significant beyond the .05 confidence limit. The homogeneous keys showed significantly less shrinkage than the empirical keys beyond the .01 confidence level on military grade and pass/fail, and beyond the .05 confidence level on academic grade.³

PSYCHOLOGICAL COMPARISON OF THE KEYS

The last comparison to be made between the empirical and the homogeneous keys was the degree to which the scores on each led to a better understanding of the criteria, that is, the degree to which each was psychologically meaningful. In a pamphlet prepared to set forth the objectives of the CCS curriculum, certain traits were hypothesized which seemed to discriminate the superior officer from the poor officer (25). With these desirable traits as the criterion, three judges, independently and later jointly, examined the keys a posteriori from the standpoint of better criterion definition.

After an examination of the items of the empirical keys, there was unanimous agreement that about two-thirds of the items bore no logical relationship with the criterion and that most of the remaining items bore only indirect relationship at best. Examples of these items which were found in two or more of the empirical keys with a positive validity are, "played card games in childhood," "carried on woodworking and cabinet-making as a hobby," and the items with significant negative validity were, "having ridden a horse in childhood" and "having driven a motor boat."

From another point of view, at least two desirable traits of the superior officers, which the judges agreed were measured by various items in CE608C, were superior scholarship and cooperation with fellow workers or group participation. Examination of the valid items hypothetically related

³ Since the standard error of shrinkage on pass/fail is based upon the transmutation of biserial r to Fisher z , it is probably an underestimate. Caution should be exercised in the interpretation of the overestimated critical ratio.

Table 17

Cross-Validation of Empirical and Homogeneous Keys

(Sample: officer candidates of Classes 51-A and 51-B)

<u>Criterion</u>	<u>Key</u>	<u>N</u>	<u>Mean</u>	<u>SD</u>	<u>Cross- validity r</u>	<u>C.R. of differences of cross- validities</u>
Final grade	Empirical	306	18.32	3.02	.22	.90
	Homogeneous	302	5.19	.94	.15	
Military grade	Empirical	306	27.57	3.24	.17	1.16
	Homogeneous	302	5.15	.68	.26	
Academic grade	Empirical	306	23.19	3.23	.30	1.83
	Homogeneous	302	5.14	.72	.16	
Pass/fail	Empirical	337	8.58	1.74	.17 ^a	.64
	Homogeneous	332	.85	.14	.22	

^a Biserial correlation coefficient, where $p = .91$ and $q = .09$.

Table 18

Comparison of the Shrinkages of the Empirical
And Homogeneous Keys After Cross-Validation

Criterion	Key	Validation r _a	Cross- validation r	Shrinkage ^b	C.R. of shrinkages	C.R. of difference of shrinkages
Final grade	Empirical	.43	.22	.24	3.09 ^e	1.23
	Homogeneous	.28	.15	.14	1.76	
Military grade	Empirical	.54	.17	.42	5.46 ^e	3.99 ^e
	Homogeneous	.35	.26	.10	1.22	
Academic grade	Empirical	.58	.30	.35	4.52 ^e	2.10 ^d
	Homogeneous	.33	.16	.18	2.35 ^d	
Pass/fail	Empirical	.51 ^c	.17 ^c	.39	5.03 ^e	4.00 ^e
	Homogeneous	.29	.22	.08	.97	

^a Correlations obtained by the validation of the homogeneous keys are multiple R's.

^b Shrinkages were obtained by appropriate Fisher z transformation.

^c Biserial correlation coefficient, where $p = .91$ and $q = .09$.

^d Significant beyond the .05 level of confidence.

^e Significant beyond the .01 level of confidence.

to scholarship showed that, as expected, the subjects having superior high school grades excelled in CCS. Other items are keyed as valid, however, which on the surface appear inconsistent with superior scholarship, e.g., the more successful officer candidates had only a high school education or less. Whatever were the conditions which caused the lesser educated subjects to excel in CCS, it is logical to assume that from year to year such conditions would not be repetitive. Two items which were related to scholarship but were keyed in a direction contrary to expectations, include: (1) adverse feelings toward education, and (2) an 8th grade education or less for fathers of officer candidates.

The second hypothesis, cooperative working with others, which also seemed to be measured by CE608C, was concerned with several items keyed as valid which would appear to discriminate the more cooperative from the less cooperative. However, other items were keyed as valid which appeared oppositely related to social cooperation and participation, such as preference for working alone and no experience as an instructor or group leader or desire to be one. At the same time such items as the desire to advise or help others and active participation in various club activities remained unkeyed. The attempt to isolate the above two "desirable" traits by examination of the valid items was fruitless.

An equally critical appraisal should be made of the psychological meaningfulness of the homogeneous keys with the objective of a better understanding of the criteria. On the basis of an inspection of the item content, descriptions of the 13 categories are given below. The descriptions "typify" the high-scoring individual.

1. Mechanical Aptitude: A person scoring high in this category has carried on woodworking as a hobby, has a shop in the home, has excelled in shop work in school, and he has made various kinds of mechanical repairs in his youth as well as in adulthood.

2. Athletic Experience: This individual has engaged in various team sports, often as a captain or coach. He has frequently engaged in various types of individual sports, and he has excelled in physical training in school.

3. Childhood Games: This subject, as a child, has participated in such games as playing checkers, dominos, and card games, digging caves, and building club houses.

4. Playboy: This person has participated in various forms of gambling in high school. He prefers playing poker over playing softball, winning a large sum of money over finding a similar unclaimed sum, working from 9:30 to 5:30 over 7:30 to 3:30, a clever friend over an honest one, and staying at home to read over going on a hike. He will not believe in or is unable to stick to a budget, and he will frequently go nightclubbing during recreational hours.

5. Socio-Economic: In the home of the high-scoring subject of this category there would be such things as a waffle iron, vacuum cleaner, extension telephone, television set, automatic water heater, and a large number of books. The father and mother of this subject have at least entered high school, and the subject has no more than two siblings.

6. Schizoid: This individual doesn't like to talk over personal problems. He doesn't expect his friends to help him out of a jam. He feels that what other people do is their business, and he prefers to be left alone. He has few friends, if any.

7. Parental Criticism: This high-scoring subject has often been criticized by his parents over such issues as relations with the opposite sex, gambling, smoking, drinking, choice of career, and not attending church.

8. Extroversion: This person has been a leader in school or a club, a class officer, debater, active member in dramatics, an instructor, and/or a camp leader.

9. Itinerant: This individual has hitch-hiked farther than 100 miles on a trip before completing high school. He prefers work with opportunity for travel and adventure over good pay and promotion, working in different places over working in the same building, changing jobs often over working at the same job, being sent overseas over staying in the United States.

10. Scholarship: This person has excelled in all courses in high school; he has never failed a course. He has often visited a library or museum in his recreational hours or on vacations.

11. Societal Acceptance: This subject believes that laws, judges, and juries are not prejudicial, that there is much fun and few worries in life, and that education does not lead to discontent. He further is against crossing picket lines and is in favor of labor's striking. He would also not prefer more color in the Air Force uniforms.

12. Childhood Responsibility: Prior to high school this subject rode an interurban bus or train alone. He has had the responsibility for the care of a pet. He has used a charge account and has owned a car when in high school, and has made a business deal in excess of \$500.

13. Aggressiveness: This is the general factor. This high-scoring individual has had fist fights in his youth. He also gambled and made long-distance calls before he was 18 years old. He was very athletic, having captained or coached a team. He has been fairly proficient in such sports as diving, boxing, wrestling, and football. He admits beating someone in a trade, and having taken advantage of someone slyly. He has been the leader of public meetings and bull sessions, and engages or has engaged in many dates per week.

It may be noted from the above descriptions that, in contrast to the empirical keys, all the categories deal with a central theme of greater or lesser complexity. The comparison of the two sets of keys as they relate to criterion definition is discussed in the next section.

INTERPRETATION OF RESULTS

Evaluation of the Cross-Validation

With reference to the hypotheses previously stated, the following conclusions are indicated by the data of this study, and each is discussed briefly in turn:

1. The empirical keys contained higher correlations with the criteria than the homogeneous keys on the development sample. As it was previously stated, Biographical Inventory CE608C was originally devised by the selection of valid items, and item inclusion in the empirical keys was based upon the unique contribution to that validity. In addition, it was found that 40 to 46 per cent of the items constituting the empirical keys were either too heterogeneous or not in sufficient number to be included in the homogeneous keys. This represented a considerable source of validity untapped by the homogeneous keys.

2. The shrinkages of the empirical keys were significantly greater than the homogeneous keys. Since the empirical keys had higher correlations with all criteria, greater shrinkage might be related to a larger original correlation rather than or in addition to the differences in homogeneity. A research design to discover these relationships would require the comparison of shrinkages of a large number of both homogeneous and heterogeneous keys. This laborious job is beyond the scope of this study.

3. Neither method of keying yielded superior validities. While the difference between the validities was not significant, the empirical keys yielded higher validities for the prediction of academic grades and final grades, and the homogeneous keys yielded higher validities for the prediction of military grades and pass/fail. This seems worthy of further investigation, since it is possible that empirical keys may relate to the prediction areas already accounted for by aptitude and achievement tests, while homogeneous keys may relate to the relatively unexplained social area.

4. The homogeneous keys were psychologically meaningful while the empirical keys were not. Among the objectives of keying a heterogeneous test should be included not only the prediction of the criterion but also the increased understanding characteristic of most criteria. Increased knowledge of the criterion will help to give a clearer perspective for the development and execution of a training program and a clearer picture of the actual versus the probable measures of success. The extent to

which the two methods of keying have added to knowledge of the criterion should be examined critically.

It was noted how inadequate the empirical keys were in criterion definition. Only about one-third of the items in the empirical keys could be indirectly related to desirable traits of superior officers, as set forth by command judgment. Since the items comprising the empirical keys were each equally weighted unity, it was impossible to know which factors to invoke to explain the criterion variance accounted for by the key.

In contrast to the empirical keys, each of the homogeneous keys were relatively easy to define. The part that each key played in explaining the criterion was indicated by its beta weight in a multiple regression equation. Coincident with criterion definition, the test constructor is given many clues as to how the multiple correlation may be increased by the addition of any missing homogeneous tests and by increasing the breadth of the more relevant scales.

It should be pointed out, however, that the validities, to which the discussion of criterion definition has been relevant, ranged from .15 to .30. The insights into the criteria which are provided by the keys cannot be related, therefore, to more than 2 to 9 per cent of the criterion variance. It must be concluded that the greater utility which is posited for the homogeneous keys is based on intuitive and not empirical grounds.

Empirical Versus Homogeneous Keying in a Program of Research

The last comparison to be made between the homogeneous and empirical keys is the manner in which both keys fit into an extended program of research. Since a good deal of time and effort is usually expended in order to evolve fairly stable keys, the job of keying is usually carried on with the purpose of long-range use. It should be noted, particularly with biographical or attitudinal-type information, that periodic re-validation of the items is essential. Items relating to socio-economic areas, educational areas, and broad attitudinal questions concerning personal adjustment are just a few types of items containing transient validities, both from time to time and from group to group. Anastasi (2) states that the distinction between the test and the criterion is merely one of practical convenience, and she urges that every test score be operationally defined in terms of empirically demonstrated behavior. The literature is replete with the many ways by which criteria may be biased (cf. Brogden and Taylor 3). Validation of the items must, therefore, keep pace with the vagaries of criterion change, and it is in this regard that the question should be asked, "how difficult would it be to keep each set of keys up to date?"

Unless only slight changes occur either in the revision of the criteria or in the inclusion of additional items, empirical keying would have to start entirely anew. A priori analysis is usually too gross to estimate

accurately how "slight" the changes are in the criterion from year to year, and which items are most affected by such changes. In addition, it is apparent that with the appearance of each new criterion, a new keying procedure would be required. On the other hand, insofar as the homogeneous keys are concerned, the entire keying procedure would have to be repeated only with very gross changes in the test itself. Where there were either revisions of the criteria or additions of new criteria, the same homogeneous keys could be used to obtain new series of significant beta weights. This procedure involves nothing more than re-validating each key on each new criterion and computation of the multiple-regression coefficient. Where additional homogeneous tests are to be devised to measure inadequately covered areas of the criterion, the old homogeneous categories can be retained, and the statistical labor of category evolution and refinement need only be concerned with the new categories. It may be seen clearly that homogeneous keying, in contrast to empirical keying, is amenable to an expanding and continuous research program.

SUMMARY AND CONCLUSIONS

This study utilized two different approaches in the selection and weighting of items for the prediction of an external criterion. The first or empirical approach has been and is today more commonly used in the construction of scoring keys. In this method the behavior to be predicted was predefined by means of an objective criterion external to the group of items which would later constitute the test. The second or rational approach developed with the lack of suitable external criteria. It was noted that even though suitable criteria were nonexistent, certain rational hypotheses about the behavior to be predicted might be agreed upon by experts, and items then written to measure such behavior. The value of each item would then be determined by the extent to which it measured the behavioral complex that the entire test measured.

It is apparent that, in contrast to the empirical method, the selection of items on the basis of internal consistency would result in a test of narrow significance in relation to the criterion, especially in the case where the behavior to be predicted was itself poorly defined. Realizing this inadequacy, test makers then resorted to the use of unrelated groups of rational hypotheses and the consequent construction of multiple tests, each of which was to represent a portion of the criterion complex.

In this study the latter approach has been somewhat departed from inasmuch as the study was restricted to the use of the previously constructed Biographical Inventory CE608C. This inventory grew out of the compilation of the most valid items of previous inventories plus additionally edited items, and it was used experimentally by the United States Air Force. Even though the inventory was not developed in accordance with predetermined rational hypotheses, fortunately, it was later shown that the items could

be analyzed into meaningful subgroups. It was thus possible to analyze the Biographical Inventory with both the rational and empirical approach. The study was designed in order to be able to key this heterogeneous assortment of biographical items systematically by the two independent methods and to follow with a statistical and psychological comparison, including the validation of the rational or homogeneous keys and a cross-validation of both sets of keys on a subsequent sample.

The sample with which both keys were validated and cross-validated was officer candidates in the Air Force. Since there was an insufficient number of officer candidates who had been administered CE608C and for whom criterion grades were available, the homogeneous keys were developed on a sample of 1000 basic airmen from the airman population.

The homogeneous keys were derived by the method of maximizing test saturation. This method basically maximizes the item contribution of common factor variance to the total variance of the test. Out of 183 items available for keying, 111, or 58 per cent, were used to evolve 12 fairly independent homogeneous categories (average $r = .15$). Seven items unused in the independent categories plus eight items which were used in the independent categories were combined to form a thirteenth category. This category correlated high with one-half of the independent categories and thus tended to be a general factor.

By the Gleser-DuBois method for maximizing test validity four empirical keys were developed on the four criteria: final grade, military grade, academic grade, and pass/fail. The keys were composed of 39, 40, 39, and 19 items or 20, 21, 20, and 10 per cent, respectively, of 183 items available.

The empirical keys yielded four correlations with the criterion for the sample on which they were constructed, ranging from .43 to .58. Validation of the homogeneous keys on the same sample resulted in four multiple correlations ranging from .28 to .35. The independence of the homogeneous keys, excluding the general factor, held up in this sample since the average intercorrelation increased less than .01.

The cross-validation of both sets of keys on an external sample resulted in considerable shrinkage which may have been caused by criterion instability, or by the capitalization on chance error in the first sample. The cross-validity coefficients ranged from .17 to .30 for the empirical keys and from .15 to .26 for the homogeneous keys.

On the basis of the statistical and psychological comparisons made between the two sets of keys, the following conclusions are drawn:

1. While few homogeneous key validities were significant, the multiple correlations of the optimally weighted keys against each criterion were highly significant. This was caused by the fact that the valid variance of the individual keys was fairly specific.

2. The independent homogeneous keys accounted for most of the valid variance in each multiple correlation; therefore, the homogeneous key resembling a general factor added negligibly to the multiple.

3. Both the empirical and homogeneous keys yielded significant validities.

4. A comparison of the validities indicates that neither method of keying proved superior.

5. Both sets of keys showed significant shrinkages, with the empirical keys showing significantly greater shrinkage for all four criteria than the homogeneous keys. This can be explained by the greater capitalization on chance error by the empirical method.

6. The homogeneous keys were psychologically meaningful and the empirical keys were not. The former should therefore provide more clues for criterion definition and revision; however, the validities of this study were of insufficient magnitude to demonstrate this empirically.

On the basis of the above conclusions and within the limitations of this study, it is recommended that where a heterogeneous test is being keyed on strictly an empirical basis, the method should be evaluated in relation to criterion improvement and understanding as well as prediction.

BIBLIOGRAPHY

1. ADKINS, DOROTHY A., and TOOPS, H.A. Simplified formulas for item selection and construction. Psychometrika, 1937, 2, 165-171.
2. ANASTASI, ANNE. The concept of validity in the interpretation of test scores. Educ. psychol. Measmt., 1950, 10, 67-78.
3. BROGLEN, H.E., and TAYLOR, E.K. The theory and classification of criterion bias. Educ. psychol. Measmt., 1950, 10, 159-186.
4. BURCS, O.K. (Ed.) The third mental measurements yearbook. New Brunswick: Rutgers Univer. Press, 1949.
5. CRONBACH, L.J., and DAMRIN, DORA E. How to determine and interpret test homogeneity. Paper read at Midwest. Psychol. Ass., Detroit, May, 1950.
6. DuBOIS, P.H., LCEVINGER, JANE, and GIESER, GOLDINE C. The construction of homogeneous keys for a biographical inventory. San Antonio, Tex.: Human Resources Research Center, Lackland Air Force Base, May 1952. (Research Bulletin 52-18.)

7. FLANAGAN, J.C. A short method for selecting the best combination of test items for a particular purpose. Psychol. Bull., 1936, 33, 635-666.
8. GIESER, GOLDINE, C., and DuBOIS, P.H. Successive approximation method of maximizing test validity. Psychometrika, 1951, 16, 129-139.
9. GUILFORD, J.P. Factor analysis in a test development program. Psychol. Rev., 1948, 55, 79-94.
10. GUILFORD, J.P., and LACEY, J.I. (Eds.) Printed classification tests. AAF Aviation Psychology Program Research Report No. 5, 1948.
11. GULLIKSEN, H. Theory of mental tests. New York: John Wiley and Sons, Inc. 1950.
12. HORST, P. Item selection by the method of successive residuals. J. exp. Educ. 1934, 2, 254-263.
13. HORST, P. Item selection by means of a maximizing function. Psychometrika, 1936, 1, 229-244.
14. LECZNAR, W.B. Evaluation of a new technique for keying biographical inventories empirically. San Antonio, Tex.: Human Resources Research Center, Lackland Air Force Base, March 1951. (Research Bulletin 51-2.)
15. IENTZ, T.F. Evaluation of methods of selecting test items. J. educ. Psychol., 1932, 23, 344-350.
16. LCEVINGER, JANE. A systematic approach to the construction and evaluation of tests of ability. Psychol. Monogr., 1947, 61, No. 4.
17. LCEVINGER, JANE. The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. Psychol. Bull., 1948, 45, 507-529.
18. LONG, J.A., and SANDIFORD, P. The validation of test items. Univer. of Toronto: Dept. of educ. Res.. Bull. No. 3, 1935.
19. PINTNER, R., and FORLANO, G.A. A comparison of methods of item selection for a personality test. J. appl. Psychol. 1937, 21, 643-652.
20. RICHARDSON, M.W. The relation of difficulty to the differential validity of a test. Psychometrika, 1936, 1, 33-49.
21. RICHARDSON, M.W. Notes on the rationale of item analysis. Psychometrika, 1936, 1, 69-76.

22. RICHARDSON, M.W., and ADKINS, DOROTHY C. A rapid method of selecting test items. J. educ. Psychol., 1938, 29, 547-552.
23. SWINNFORD, F. Validity of test items. J. educ. Psychol., 1938, 27, 68-72.
24. TOOPS, H.A. The L-method. Psychometrika, 1941, 6, 249-266.
25. USAF Officer Candidate School, Lackland Air Force Base, Special Order 14, June 1950.
26. RUBIN, J. The method of internal consistency for selecting test items. J. educ. Psychol., 1934, 25, 345-356.

Manuscript received 29 September 1952.