

THE AMERICAN UNIVERSITY LIBRARY

Manuscript Theses

Unpublished theses submitted for the Master's degree and Doctor's degree and deposited in The American University Library are open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but passages may be copied only with permission of the author, and proper credit must be given in subsequent written or published work.

This thesis by _____ has been used by the following persons, whose signatures attest their acceptance of the above restrictions.

A library borrowing this thesis for use by its patrons is expected to secure the signatures of the users.

Ronald H. Gubb

NAME AND ADDRESS

DATE

Ronald H. Gubb 5605 Monwest, Hyattsville, Md. 2-26-54
Howard W. Garlin 7604 Gresham St. Springfield, Va 12-9-61

Bryant, Bernice Morgan, 1908-

George Gershwin, young composer, by Bernice Bryant.
Illustrated by Nathan Goldstein. Indianapolis, Bobbs-Merrill [1965]

200 p. col. illus. 20 cm. (Childhood of famous Americans)

Bibliography: p. 198.

1. Gershwin, George, 1898-1937--Juvenile fiction.

PZ7.B834Ge

65-23666

Library of Congress

[18-1]

15355

OBJECTIVE SELECTION OF SUPERVISORY PERSONNEL
IN PUBLIC MANAGEMENT

by

Joseph G. Colmen

Submitted to the
Graduate Faculty of The American University
in Partial Fulfillment of
the Requirements for the Degree
of
Doctor of Philosophy

Signatures of Committee:

Chairman: _____

ACKNOWLEDGEMENTS

The author wishes to express appreciation for the generous support given by Mr. John A. Watts, Director of Civilian Personnel, and Mr. Jack Pockrass, Chief of the Placement and Employee Relations Division, both in Headquarters, U. S. Air Force, under whose cognizance the project described in this dissertation was completed. The professional assistance of Mr. G. O. Fiedler, Mr. James H. Blackburn, Mr. Laurence Kashdan, Mr. Charles A. Scoggins and Mrs. Ruth Raffaelli are gratefully acknowledged. Permission given to the author to use or adapt certain test material developed by the U. S. Civil Service Commission is also appreciated.

The opinions expressed herein are entirely those of the author and do not in any way reflect the policies or official views of the Department of the Air Force.

TABLE OF CONTENTS

CHAPTER	PAGE
I. PURPOSE OF THE STUDY	1
II. REVIEW OF THE LITERATURE	6
Early studies	8
Descriptive studies	20
Supervisory selection programs based on experience or opinion	22
Interest measures	26
Attitude measures	29
Biographical Information Blanks	31
Measuring administrative judgment	33
Reading ability	36
Ratings as predictors: sociometric and sociodramatic methods	36
Ratings as predictors: additional methods	39
Personality measures	41
General studies	45
Supervision defined	49
III. THE CRITERION MEASURES	74
Ingredients of success as an Air Force supervisor	75
Precautions to be taken in obtaining ratings	80
Method of obtaining superiors' ratings.	85
Method of obtaining subordinates' ratings	88

CHAPTER	PAGE
Distribution of ratings	90
Reliability of ratings	92
Combining the ratings	94
IV. THE SAMPLE	98
Type of sample	100
Selection of bases	100
Selection within bases	102
Criteria for inclusion in sample	103
Characteristics of the sample	105
V. THE EXPERIMENTAL BATTERY	106
Description of tests	110
The test session	118
VI. ANALYSIS OF DATA	121
Design of analysis	121
Keying personality tests	125
Steps leading to most predictive final personality battery	130
The aptitude tests	147
The final overall battery	149
Availability of tests in the final battery	153
VII. PROBLEMS IN USE OF THE BATTERY	155
Standardized directions	156
Directions for interpretation	157
Use of non-test variables	158
Local administrative determinations	158
Public relations considerations	161

CHAPTER	PAGE
Security of test materials	169
Providing for economy of operation	170
Practical problems of research in an administrative setting	172
VIII. SUMMARY AND CONCLUSIONS	174
Summary	174
Conclusions	178
Suggestions for further research	180
BIBLIOGRAPHY	183
Appendix	192

LIST OF TABLES

TABLE	PAGE
I. Ability of Three Tests to Predict the Criterion of Alternative Ranking . . .	88
II. Distributions of Ratings Completed by Superiors at Both Installations . . .	91
III. Distributions of Ratings Completed by Subordinates at Both Installations . .	91
IV. Reliability of Superior and Subordinate Ratings by Base	93
V. Population Statistics	101
VI. Occupational Distribution of Supervisors in the Sample by Base	102
VII. Tests Used in the Experimental Battery .	109
VIII. Correlations of the Personality Sub-Tests with the Criterion	132
IX. Correlations of Personality Tests with the Criterion	137
X. Intercorrelations of Tests in Final Personality Battery in terms of Frequency	144
XI. Correlations of Tests in Final Personality Battery with the Criterion . . .	145
XII. Comparison between Air Force Bases in Scores on the Personality Tests . . .	145
XIII. Number of Responses Keyed in Final Personality Battery	147
XIV. Matrix of Intercorrelations of Final Battery Tests and Combined Criterion .	150
XV. Increase in Predictive Efficiency by Use of Proposed Testing Battery . . .	163
XVI. Expectancy Chart for Proposed Battery .	166
XVII. Expectancy Chart for Proposed Test Battery Based on Extreme Groups Only . .	167

CHAPTER I

PURPOSE OF THE STUDY

American corporate enterprise has reached gigantic proportions and stands today as testimony to its own successful operations. The advances made by inventive researchers and engineers have contributed greatly to a corresponding growth of the vast business machine. The productivity of the American worker has also played its part in this development. The foresighted investment of capital likewise enjoyed a leading role in the drama of industrial maturation. The application of progressive management techniques and the emphasis of improved management-employee relations no doubt created the climate in which this venture was able to flourish.

In comparison, the proper selection and placement of manpower may seem a somewhat inconspicuous chore. And yet it may be reasoned, that the industrial advances of the world were made in spite of the ability to apply techniques of selection that we know now contribute so directly to the efficiency of labor and to the productivity of organizations.

When they became available, however, industry was fast to make direct use of the results of psychologists' studies into the area of selection. And industry soon learned of the advantages to be derived from better selection in terms of increased productivity, reduced turnover and absenteeism, reduced training time requirements, improved morale and fewer grievances. But somehow industry's greatest need in the selection area was either never posed as a problem for the psychologist to investigate or the psychologist's fund of techniques and methodology had not advanced sufficiently for him to tackle the problem. Thus, research into selection of supervisors was, with noteworthy exception, put off until less than a decade ago. The psychologist's storehouse of tools of measurement were increasing in many other occupational areas, however, such as those developed for selection of street railway operators, clerical workers, department store salespersons, and the like. Nevertheless, as late as 1951, Castle and de la Garforth concluded from an extensive review of the literature that adequate valid procedures for selection of supervisors did not exist, due probably to the lack of a satisfactory

criterion of success.¹

Whether or not industry recognized the supervisor as the key to organizational efficiency, shopworn methods continued in use for his selection, even as improvements were being wrought in the selection of his subordinates. The folly of the situation may be readily understood when one stops to consider the effect of a poor supervisor on a presumably capable worker. Nonetheless, supervisors continued to be drawn from the ranks or the outside on the basis of such factors as seniority alone, technical proficiency alone, nepotism alone, or any combination of these factors. Even today within government, as Melvin Purvis has recently observed, these same anachronistic techniques are being used in most of the departments of the huge federal establishment.²

¹P.F.C. Castle and F. de la Garforth, "Selection, Training and Status of Supervisors: I. Selection," Occupational Psychology, 25:109-123, April, 1951.

²Senate Report No. 2100, 82nd Congress, Supervisory Selection in the Federal Government. Washington, D.C., U. S. Government Printing Office, 1952, p. 2.

The present study attempts to investigate intensively the possibility of developing valid objective selection techniques for the selection of supervisors within the Department of the Air Force. Stated differently, it is the hypothesis of this study that discrete series of behavioral components exist which can be operationally defined and perceptually and verbally described as supervisory ability, and that this ability can be measured by appropriate scales.

The dissertation is organized so that there will first be an historical discussion of past efforts at objective selection of supervisors. Supervision will be defined. The criterion measure against which tests will be evaluated as predictors will be described and will be followed by a description of the experimental tests used. The data analysis procedure will be amply detailed and research results will be summarized. Procedures for collection of criterion data, selecting the sample, conducting the test sessions and processing of the raw data are described in their appropriate chapters.

In addition, this dissertation will include a survey of administrative considerations in implementing the resultant test battery with a minimum of disruption to existing personnel operations at numerous and diverse field offices.

CHAPTER II

REVIEW OF THE LITERATURE

A survey of the literature on selection of supervisors reveals several interesting and significant points. First, few attempts to objectify selection of supervisors were reported before 1940. Several of those experimental attempts seem modern by contrast, however, in terms of recognition of the types of measuring devices to be used, criterion data employed, and statistical techniques utilized. These earlier studies will therefore be elaborated upon below. Second, much space has been allotted in a variety of publications to what may be referred to as "armchair" analysis of the supervisory job and development of proposed traits or sets of predictors. Except to mention a few, relatively little attention will be devoted to this more abundant, but quite sterile portion of the historical literature. Third, many experimental studies which have been reported have been open to attack on the basis of employing inadequate research design; utilizing experimental tests with little logical foundation for their selection; accepting samples too small for

drawing conclusions; omitting entirely or collecting sketchy or unreliable criterion data; applying improper statistical techniques; or drawing unwarranted conclusions from the results. More attention will be devoted below to these studies in an attempt to demonstrate the pitfalls which should be avoided in a research project of this kind and which, it is hoped, have been avoided as much as possible in this study.

One additional curiosity in the literature which continued into present day research is the apparent lack of recognition of the difference which exists between executive, administrator and leader on the one hand, and supervisor on the other. The terms are most often used interchangeably and infrequently do authors attempt to define the specific group to which they refer in their reports. Because of this confusion, the following review includes references to research studies on executives and administrators in the hope that some light may be shed on methodology which may be promising in improving selection of supervisors.

I. EARLY STUDIES

One of the earliest attempts at discovering a predictor of executive success was reported by Bingham and Davis¹ in 1924. They applied an intelligence test and found that alone it was insufficient as a means of predicting success. This finding, while negative in application, did suggest the complexity of the occupational area being studied and the need for measuring other, perhaps as yet more obscure psychological characteristics responsible for success in the executive job.

In 1925, Bills, using the case study approach, uncovered a fact which today sounds somewhat commonplace but which then must have made the task of selection of managers seem even more impossible.² Bills analyzed the jobs of two store managers as well as the incumbents in them. One of the managers Bills found to be quicker mentally, the other more astute in social relations, yet both made out equally well enough to justify retention.

¹Walter V. Bingham and W.T. Davis, "Intelligence Test Scores and Business Success," Journal of Applied Psychology, 8:1-22, March, 1924.

²Marion A. Bills, "Predicting Managerial Success: A Case Study of Two Businessmen," Journal of Personnel Research, 4:46-51, June, 1925.

Bills concluded that

in reality the method of doing a complex job may differ so radically from person to person that the job actually becomes different, and consequently permits men of different qualities to fill it. That is, the psychologist cannot hope for a correlation between tests and success if the actions involved in the job for which one is testing are not standardized.³

While this conclusion, based on two cases, has a great deal of apparent truth in it, there seems to be a core of related or similar demands made of all incumbents in supervisory or executive positions to make possible the predictive measurement of qualities. The extent to which individual positions differ, however, may account in part for the reductions one obtains in validity coefficients.

Five years later, Beckham and Levine reported a study on the selection of executives in the Cincinnati Civil Service Commission, utilizing for the first time a battery of tests, presumably to enable them to measure those qualities over and above the intellectual factor which, as reported above, had been found insufficient

³Ibid., p. 46.

for measuring executive success.⁴ For the experimental battery, Beckham and Levine selected (1) The G.W. and F. H. Allport A-S Reaction Study, to measure ascendancy or self-assertion versus submission, (2) The Laird Personal Inventory C-2 Study of Colgate University Laboratory, to measure introversion-extroversion, and (3) a test of following written instructions. The authors of this study appear to have recognized the need for a measuring stick in the personality sphere, based possibly on a more intent look at the several demands in the personality area made upon incumbents of executive positions. Twenty-nine executives were included in the study, along with thirty-one water meter readers used as a control group. For a measuring rod of the success of executives in the sample, efficiency ratings for the past year's performance were used. The ratings on those executives also rated by the City Manager were found to have a reliability of .94. The validities were as follows:

	Execu- tives	Water Meter Readers
r (Allport-A-S versus Efficiency Rating)	.25	.21
r (Laird Personal Inventory vs. Rating)	.01	-.07

⁴R. O. Beckham and Michael Levine, "Selecting Executives," Personnel Journal, 8:415-420, April, 1930.

The authors concluded " . . . an executive should be ascendant as tested by the Allport and Allport A-S Reaction Study, and . . . it does not matter whether he is an introvert or extrovert as indicated by the shorter form of the Laird Personal Inventory C-2 Test.⁵ Insofar as the following directions test was concerned, no correlations were reported, but the authors suggest that executives should be higher in score than workers they supervise on this test. While this study is deserving of praise for its attempt at investigating the personality component of the executive job, the small number in the sample and generally unsatisfactory nature of Civil Service efficiency ratings as research criteria throw some doubt on the meaningfulness of the findings. In addition, no attempt appears to have been made to include in the statistical analysis all test data so that one can conclude little about either the interrelationships between the tests or their predictive powers as a multiple battery.

Possibly one of the early studies which appears

⁵Ibid., p. 420.

most modern in comparison with present day research reports in this field is one reported by Uhrbrock and Richardson in 1933.⁶ Using 163 supervisors responsible for directing 2100 factory workers in a large industrial organization, the authors administered a comprehensive test battery which included the following:

- (1) Modified Army Alpha (Bureau Test VI)
- (2) Strong's Vocational Interest Blank
- (3) McCall's Multi-Mental Test
- (4) Minnesota Paper Form Board (Series A)
- (5) Minnesota Paper Form Board (Series B)
- (6) Selected items from Thurston's Personality Schedule
- (7) Multiple-choice, Company Information Test
- (8) True-false, Company Information Test
- (9) Completion form, Company Information Test

These represented 820 items, in addition to which a personal history record was included. For a yardstick of success, Uhrbrock and Richardson used order of merit ratings, paired comparison ratings and graphic rating scales which requested judgments on such qualities as flexibility, quality and quantity of work, planning ability, analytical ability, leadership, cooperativeness, teaching ability, and the like. Each judge or rater rated forty-five men on all three types of ratings. The results indicated that only 85 of the original 820 psychological and interest

⁶R.S. Uhrbrock and Marion W. Richardson, "Item Analysis: The Basis for Constructing a Test for Forecasting Supervisory Ability," Personnel Journal, 12:141-154, July, 1933.

items were significant discriminators. Company information test items had proportionately greater numbers of valid items than purchased tests. The correlation between the 85 items and the ratings was .71. Four personal history items significantly differentiated between men in the poorest third and men in the upper two-thirds of the group of supervisors. They were: (1) age, (2), schooling, (3) self-evaluation of blueprint reading ability, and (4) military service record. A physical examination total score correlated with the criterion .12, no single item in the examination alone being significant. So far as can be determined, this study represents the first in which biographical information items were used in addition to personality, interest, information and intelligence test items. The mortality rate of the items (735 out of 820) is even today one of the most irritating problems in the measurement of personality, interest and attitudes. The correlation of .71 reported for these items smacks of being a "boot-strapped" correlation, that is, based on items which were pre-selected because of their known relationship to the criterion, then grouped and correlated statistically with the criterion. It does not appear that the authors attempted to cross-validate

the items on an entirely independent sample to see whether in fact, the items retained their predictive value. This criticism, it will be seen however, may be leveled against more modern researchers for whom there is less excuse than for the writers of 1933.

In Great Britain in 1938, Mitchell reported a study in selection of store managers.⁷ In a preliminary study, Mitchell classified fifty men in two groups: (1) the "field" type of manager, and (2) the "office" type of manager. Next he obtained detailed personal history items from 183 managers. The more successful managers were on the average fourteen pounds heavier in weight than the less successful. Mitchell hypothesizes that

The weight differences, together with general observations made of the physical and mental characteristics, suggest that the group of more successful managers, the 'field' type, include a large proportion of 'cyclothymes' ('pyknic' type) and that the less successful group, the 'office' type, contain a large proportion of 'schizothymes' ('aesthenic' or 'athletic' type).

⁷J. H. Mitchell, "An Experiment in the Selection of Sales Managers," Occupational Psychology, 12:308-318, Autumn, 1938.

⁸Ibid., p. 311.

Mitchell went further in this study, applying a battery of tests to the selection of sales managers. As a criterion, he required an ability in the tests to differentiate between district sales managers and salesmen. This of course poses an entirely different problem. The ability of a test to differentiate between occupational groups is far different from its ability to select potentially successful workers in any given occupation, from among all applicants for that position. Nonetheless, Mitchell used the following battery:

- (1) NIIP Group Test 33 of general mental ability
- (2) Vocabulary Test
- (3) Beckman's revision of the Allport and Allport A-S Reaction Test
- (4) Surgency-Desurgency Test, which included
 - (a) A word-association test
 - (b) Completing forms
 - (c) Word series
 - (d) Topics -- a verbal test for fluency of ideas
 - (e) Ink blots

Mitchell did not use multiple correlational analysis or the discriminant function technique in analyzing his results. His statistical approach was merely to test the significance of difference between the mean scores for each test obtained by sales managers and salesmen. The results were as follows:

Test	Salesmen N = 120 (Mean Scores)	Managers N = 42 (Mean Scores)	PE (Diff. bet. Means)
Vocabulary	85	120	3.66
Intelligence	98	119	5.87
A-S Reaction Study	33	34	2.11
Word Association	31	53	2.13
Completing Forms	8	10	.66
Word Series	32	40	1.58
Topics	325	326	18.07
Ink Blots	9	13	.82

Mitchell also supplies the verbal comment, unsupported, that superiors fell in between salesmen and managers. Using the statistics presented, it would appear that the Vocabulary test, Word Association, Word Series and Ink Blots are most able to meet the criterion of ability to differentiate salesmen from managers.

A relatively early study worth reporting because it sheds light on problems of evaluating leadership success in establishing the criterion is one conducted by Zeleny in 1939.⁹ Zeleny, in attempting to select group discussion leaders, used ratings based on voice of the candidate and observation of him; identification by status ranking, where by changing groups, everyone ranks all others; the man-to-man technique using groups of

⁹Leslie D. Zeleny, "Objective Selection of Group Leaders," Sociology and Social Research, 24:326-336, March-April, 1939.

five; and sociometric observation. Zeleny found that ratings on voice and by sight gave a fairly good, rough, preliminary rating. Identification by status ranking had a reliability of .684. On this basis, the fourteen "best" were placed in charge of groups, as were the fourteen "worst." Both were now rated and significant differences were obtained in ratings on performance of the "best" and "worst" groups. The man-to-man technique was accomplished by placing subjects randomly in groups of five. After observing performance, students then checked their choice for leader of that group. The student receiving most checks received a rank of 1, the student receiving next most checks, a rank of 2, and so on. The correlation between the ranks thus obtained and five faculty ratings was .72. It is worth cautioning that this may well represent a test of reliability rather than validity since all we know is that the students and faculty rated in the same way, not that either was necessarily a valid rating. In sociometric identification, each subject chose five students he would like to have in his group. The score was the number of times a student was chosen either by leaders or followers as determined in the earlier portion of

the study. Reliability of this technique was estimated at .94. Perhaps the most significant aspect of this study was the attempt to define leadership not only in terms of the superiors' evaluations of the student, but in terms of the subordinates' and peers' as well.

In 1940, Harrell reported an attempt to develop a test battery for cotton mill supervisors.¹⁰ Using the Bernreuter Personality Inventory, the Otis Self-Administering Examination (Higher), Moss, Hunt and Omwake's Test of Social Intelligence, Strong's Vocational Interest Blank, age and education as predictors, Harrell collected superiors' ratings on a scale of 1 to 4 for use as the criterion. Harrell recognized the probable low reliability of the ratings. The results with 42 supervisors were as follows:

¹⁰Willard Harrell, "Testing Cotton Mill Supervisors," Journal of Applied Psychology, 24:31-35, February, 1940.

<u>Test</u>	<u>Tetrachoric r</u>
Bernreuter Personality Inventory	
(Self-Confidence Key)	.15
(Sociability Key)	.09
Otis SA Examination (Higher)	.37
Moss, Hunt, Omwake's Test of Social Intelligence	.18
Age	.03
Education	.03
Strong's Vocational Interest Blank	Most interest shown in people and business occupations.

Harrell did not attempt multiple correlational analysis, so that indication of the weights of the combined tests most predictive of success as a cotton mill supervisor is not available. Evidently intelligence as a single predictor would best predict the criterion in this case, though used with the Bernreuter self-confidence key and the Moss, Hunt and Omwake Test of Social Intelligence, it is possible that this prediction could be improved. The low correlations obtained with the personality tests suggest that ready-made tests, using keys derived on totally unrelated samples can hardly be expected to provide satisfying results when applied to new populations.

In 1941, Blankenship suggested methodological approaches to the problem of measuring administrative traits.¹¹

¹¹ Albert E. Blankenship, "Methods and Problems in Measuring Administrative Traits," Public Personnel Quarterly, 2:69-72, Spring, 1941.

In actuality, his paper skirted the problem of measurement almost entirely and instead suggested means of analyzing administrative jobs or job requirements. His suggestions included the intuitional approach which merely listed traits; the biographical method; the case history approach; informal job analysis; formal job analysis; and direct analysis of administrators' characteristics, implying the use of tests in a scientific study.

The foregoing summary of work accomplished up to 1941, for all its recognized and unrecognized weaknesses, represents almost all of the advances as well as mistakes made in the succeeding years of research in the area of supervisory selection.

II. DESCRIPTIVE STUDIES

For some reason or other, the field of supervisory selection has drawn numerous writers who feel that, by describing an operational program for selection of supervisors at X or Y company, a sizeable contribution to the literature is being made. While the case study approach has merit in certain types of inquiry, in this field where evidence of validity in statistical terms is necessary to

establish a basis for a given method or set of predictors, the case study approach is much less valuable. Because of differences in techniques and instruments used in each reported case study, furthermore, it is next to impossible to draw general conclusions about the effectiveness of any. And lastly, the case study approach, by its generally optimistic tone (negative results are seldom reported) neglects to mention the necessary consideration of check validity when application is to be made under varied conditions in another organization. Because of their generally unscientific approach, therefore, little coverage will be given to them in this paper. Suffice it to mention articles by Drury,¹² Dunn,¹³ Fraser,^{14, 15} and Blake and Harrison¹⁶ as prototypes of many other similar papers which are descriptive,

¹²Lynn B. Drury, "Selecting Employees for Advancement," Personnel Journal, 20:166-171, November, 1941.

¹³Paul C. Dunn, The Selection and Training of Railroad Supervisors, Department of Business and Engineering Administration, Massachusetts Institute of Technology, 1942, pp. 3-53.

¹⁴John Munro Fraser, "An Experiment with Group Methods in the Selection of Trainees for Senior Management Positions," Occupational Psychology, 20:63-67, January, 1946.

¹⁵-----, (continued on following page)

but little more. Another study representative of many more is descriptive but boldly posits techniques for selection of supervisors without even presenting case study evidence for their validity.¹⁷

III. SUPERVISORY SELECTION PROGRAMS BASED ON OPINION OR EXPERIENCE

Several papers worth noting have inquired into qualities which appear important to success as a supervisor. While many such papers are merely the "armchair" reflections of their authors, those reported below represent a more reliable analysis, reflecting as they do the consensus of many reporters or the opinion of trained observers of psychological qualities.

Riegel reported conducting interviews with representatives of twenty companies to obtain executive opinion regarding selection and development of prospective

¹⁵(Continued from preceding page) "The Group Method of Selecting Executives," Personnel, 26:50-52, July, 1949.

¹⁶Wainwright D. Blake and Arthur E. Harriman, "The Selection and Training of Executives," Journal of Social Psychology, 29:29-33, February, 1949.

¹⁷Richard S. Schultz, "How to Develop Successful Office Supervisors," Personnel Journal, 25:273-281, February, 1947.

¹⁸ foremen. A summary analysis revealed that qualities most often sought and which were felt to relate most to success as a foreman were good health; job knowledge, skill and learning capacity; and all good personality traits, such as perseverance, initiative, dependability, ability to accept responsibility, frankness moderated with tact, self-control, cooperativeness and resourcefulness.

Gowin asked 276 businessmen and executives to rank qualities most important and least important to executive success.¹⁹ Judgment, initiative and integrity were ranked highest. Refinement, appearance and sense of humor were ranked lowest.

O'Connor reported isolated measurement of five characteristics common to one hundred Presidents and Vice-Presidents of successful companies.²⁰ They were a large

¹⁸ John W. Riegel, The Selection and Development of Prospective Foremen (Bureau of Industrial Relations, University of Michigan, Bulletin Number 11, 1941), pp. 18-19.

¹⁹ Enoch B. Gowin, Selection and Training of the Business Executive, (New York: The MacMillan Company, 1918), p. 45.

²⁰ Johnson O'Connor, Psychometrics, (Cambridge, Massachusetts: Harvard University Press, 1934), pp. 277-278.

English vocabulary; multitudinous, high-level aptitudes; objective personality; accounting aptitude; and aptitude for their first jobs.

In surveying the literature on personal factors associated with leadership, Stogdill concluded from evidence cited in at least fifteen studies that the average leader exceeds the average member of his group in intelligence, scholarship, dependability, activity and socio-economic status.²¹ Ten to fifteen studies reported that he also exceeds in sociability, initiative, persistence, knowing how to get things done, self-confidence, alertness and insight into situations, cooperativeness, popularity, adaptability and verbal facility.

Mandell reported tentative evidence for the relationship of certain tests to successful performance as a personnel worker, which is an administrative field of work.²² These were administrative judgment; vocabulary; evaluation of statements based on Roback's Mentality Test for Superior Adults in which statements must be classified as (1) striking or significant, (b) commonplace or

²¹ Ralph M. Stogdill, "Personal Factors Associated with Leadership: A Survey of the Literature," The Journal of Psychology, 25:35-71, January, 1948.

²² Milton M. Mandell, "The Selection and Promotion of a Personnel Staff," Personnel, 25:125-127, Sept., 1948.

obvious, (c) absurd, (d) tautological or (e) a ludicrous contradiction or joke; personal analysis; and the Interaction Chronograph. In spite of tentative evidence for these tests, Mandell saw the need for further study. Later Mandell offered the following factors as a basis for a selection program for supervisors:

23

- (1) People-mindedness versus stereotype-mindedness
- (2) Technical knowledge
- (3) Organizational requirements
- (4) Verbal and reading ability
- (5) Emotional stability
- (6) Knowledge of individual and social psychology
- (7) Knowledge of personnel administration and relations

These papers reveal a common thread running through analyses which may be reduced to need for certain intellectual capacities, such as judgment, reading ability and verbal ability, possibly to a degree above that possessed by their subordinates; personality characteristics associated with getting along with both superiors and subordinates; technical and administrative skills and knowledges; good health; and certain character traits such as integrity.

²³Milton M. Mandell, "Selection of Blue-Collar and White-Collar Supervisors," Personnel, 25:321-327, March, 1948.

The following sections are devoted to attempts at predicting supervisory or executive success by way of a specific type of measuring instrument, such as interest inventories, attitude measures, biographical information blanks, administrative judgment, reading ability, sociometric rating techniques, sociodramatic methods, ratings as predictors, and personality and projective devices.

IV. INTEREST MEASURES

Using 579 public administrators with a mean educational level of 15.9 years and a mean salary of \$6,000 per annum, Strong developed a scale which contrasted public administrators with men in general with respect to types of work interested in, magazines preferred, kind of people liked, activities preferred, characteristics of jobs and introspective idea of the kind of person they thought themselves to be.²⁴ Those who scored low on the scale seemed to prefer their own technical work to directing others or directing a program. In the same year, Strong reported the interests of senior and junior public administrators

²⁴Edward K. Strong, Jr., "Interests of Public Administrators," Public Personnel Review, 6:166-173, July, 1945.

differed sufficiently to suggest that one-fourth to one-third of the juniors did not have the interests of seniors and that their interests were not likely to change.²⁵ Senior public administrators had more the interests of presidents, scientists and salesmen, while junior public administrators had more the interests of social workers, production managers, general office workers, and skilled workers.

On administering his Vocational Interest Blank to 552 public administrators from a wide variety of public activities earning from \$3,000 to \$10,000 per annum, three-quarters of whom were college graduates, Strong found no A or B/ ratings on 34 occupational interest scales.²⁶ The group did score B on personnel with a score of 39; production manager with a score of 38; and lawyer with a score of 35. This suggests a tie-in with the two major functions of administrators, namely, handling of people and getting the work done. Strong concludes that the field of public administration is not clearly defined, though it may be reasoned that the practicing administrator possesses

²⁵Edward K. Strong, Jr., "Interests of Senior and Junior Public Administrators," Journal of Applied Psychology, 30:55-71, February, 1946.

²⁶"Differences in Interests among Public Administrators," Journal of Applied Psychology, 31:18-38, February, 1947.

the interests most like those he directs so that in the aggregate, the differences tend to cancel out each other.

More recently, Knaft confirmed what in essence Strong had found earlier.²⁷ Taking 38, or the top 27% and the lowest 27%, of 70 managers of bakery shops, Knaft administered the Strong Vocational Interest Blank, the Jurgensen's Classification Inventory and the Wonderlic Personnel (intelligence) Test. Analysis of thirty-nine keys on the Strong Vocational Interest Blank, using mean scores revealed that for no key, including the interest maturity key, occupational level key, or masculinity-femininity key was there a difference between the better and poorer managers which was significant at the 5% level of confidence. Knaft thereupon constructed his own bake shop manager key on the basis of items which differentiated at the 10% level of confidence or better. A cross-validation study using the new key on thirty-two manager trainees using success or failure as a criterion of job success yielded a bi-serial correlation coefficient of .53. This key correlated .27 with a special key similarly derived for the Jurgensen's Classification Inventory and

²⁷E. B. Knaft, "Vocational Interests and Managerial Success," Journal of Applied Psychology, 35:160-163, June, 1951.

.34 with the Wonderlic Personnel Test.

It seems that standardized interest inventories, particularly the Strong, if scored in accordance with pre-developed scoring keys are not likely to provide valid measures of success in managerial positions. More hope may be held out for such measures if they are keyed on the specific population for which the measure is desired, provided a system for replication of the experiment is built into the research design.

V. ATTITUDE MEASURES

The first located report on a measure of attitudes in connection with supervisory or executive success is a paper by Hersey who presented a sample self-analysis quiz which was designed to enable managerial personnel to rate themselves on agreement or disagreement with certain statements designed to reflect attitudes concerning workers and working situations.²⁸

The next reference to such a measure is found in an article by Mandell, relating supervisors' attitudes to job performance.²⁹ Mandell reported obtaining a sample of 278

²⁸Rexford Hersey, "Self-Analysis Quiz for Supervisors and Executives," Personnel, 24:454-475, May, 1948.

²⁹Milton M. Mandell, "Supervisors' Attitudes and Job Performance," Personnel, 26:182-183, November, 1949.

persons as follows: 165 first-line supervisors of skilled and unskilled trades employees; 40 first-line supervisors of professional meteorologists; and 28 first-line supervisors of clerical employees. In addition there were included 45 persons in management work, most of whom had supervisory responsibilities. The supervisory attitudes test comprised forty statements such as "most workers are less efficient today than ten years ago." Respondents were asked to check the response most nearly approximating their feelings from among "Strongly Agree," "Agree," "Undecided or Uncertain," "Disagree," or "Strongly Disagree." Mandell concluded that "poorer supervisors reveal a lack of faith in human beings - they are pessimistic and they are 'sour'."³⁰ This test had low intercorrelation with tests of supervisory judgment, reading comprehension and mechanical principles, thus apparently measuring something outside the intellectual sphere.

A test of the kind described above, in addition to its face validity, seemed to have value in measuring a component of supervisory success apart from cognitive or perceptual items. Unfortunately, correlations with a criterion were not reported, but further study in this area would appear warranted.

³⁰Ibid., p. 183.

VI. BIOGRAPHICAL INFORMATION BLANKS

In attempting to determine the extent to which certain measurable traits of group supervisors were related to success in supervising subordinates, Stockford reported obtaining rankings of subordinate supervisors from their superiors from best to poorest in terms of administrative ability, supervisory ability and technical ability.³¹ Of particular interest, though few statistical results are presented, is the application of biographical data as predictors. In this study, Stockford found that men in the Superior and Good groups had more related previous work experience than had men in the Fair or Poor group. Average length of time worked for each previous employer was significant. Superior and Good groups had taken more advanced training in related courses. Men with economic responsibilities of wife, children and home were better supervisors. In use of tests, Stockford found that personality tests yielded no valid results, but that ratings of personality by department foremen predicted success forty percent better than

³¹ Lee Stockford, "Selection of Supervisory Personnel," Personnel, 24:186-199, November, 1947.

chance. This latter fact suggests the weight given by officials responsible for selection to personality components. The California Test of Mental Maturity revealed that only twenty percent of the Superior and Good groups scored under 110, while forty-percent of the Fair and Poor groups did. There was no relationship found between seniority and success as a supervisor.

In attempting to predict proficiency of administrative personnel in schools from personal history data, Guilford and Comrey administered a 150 item, multiple-choice, Biographical Information Blank to three hundred school principals and vice-principals.³² Using promotional ratings as a criterion of success, item analysis yielded only eight predictive items, forcing the authors to conclude that the method has "limited promise of usefulness in the selection of school administrators."³³

The Biographical Information Blank is found to have had conflicting success in the two situations reported.

³²J.P. Guilford and Andrew L. Comrey, "Prediction of Proficiency of Administrative Personnel from Personal-History Data," Journal of Educational and Psychological Measurement, 8:281-296, Autumn, 1948.

³³Ibid., p. 295.

Whether or not this is due to the selection of the items, the criterion, the presence or lack of rationale behind the choice of items or the possibility of deception on the part of respondents is difficult to determine. Forced-choice techniques may help to overcome the latter objection. This area may have possibilities worthy of exploration, if items are properly developed.

VII. MEASURING ADMINISTRATIVE JUDGMENT

While judgment, intelligence and reasoning ability were early located as related to supervisory or administrative success, attempts to relate the judgment factor in tests specifically to supervisory or administrative situations did not come about until much later.³⁴ In 1947, Timpany recognized that

It was relatively easy to assess a man's or woman's technical knowledge, and, in addition, an estimate of the potential ability to acquire technical knowledge could be obtained by the use of an intelligence test. . . . But it appeared that the real problem with the remaining individuals who obtained high intelligence test scores was to decide which possessed the qualities of personality and temperament necessary in the good supervisor.³⁵

³⁴Nancy Timpany, "Assessment for Foremanship," British Journal of Psychology (General Section), 38:23-28, September, 1947.

³⁵Ibid., p. 23.

Timpany developed a test with A.W. Heim called HT1, paper-and-pencil variety, with twenty-three items, each describing a factory situation and asking what, if he were a foreman, the respondent would do in that situation. No time limit was imposed in answering the questions. The test was administered to fifty foremen in three different factories and ratings of their foremanship ability obtained in terms of ability to get along with workers, ability to get along with management, technical knowledge, organizing ability, ability to maintain discipline, and initiative and improvising ability. By calculating preference ratios for each question, a key of most acceptable responses was developed. No cross-validation was attempted so that there is no evidence that the key thus developed would retain its predictive powers on another sample.

Later, Mandell reported application of an administrative judgment test to selection of administrators, whose job he defined as one in which more than fifty percent of the time was devoted to program planning and coordination.³⁶ The test was a five-choice form with from eighty to one-hundred items. Questions were designed to measure "broad

³⁶Milton M. Mandell, "The Administrative Judgment Test," Journal of Applied Psychology, 34:145-147, June, 1950.

understanding of the process of administration; understanding of the administrative problems of large organizations . . . ; common elements in the administrative process." Mandell attempted to divorce the contents from complete dependence on basic training by "stressing problems which could be evaluated on the basis of observation and experience or training." A split-half reliability coefficient of .94 was obtained with 258 cases. To establish validity, collective ratings of colleagues and superiors as well as position grade were used. With samples ranging in size from 20 to 63, correlations with collective ratings between .49 and .68 were obtained. With samples ranging in size from 42 to 63, correlations with position grade of from .28 to .56 were obtained. Correlations from .59 to .69 of the Administrative Judgment Test with the American Council on Education Test led Mandell to conclude that the Administrative Judgment Test was not measuring the same thing, a judgment which may be open to question. Nevertheless, the tremendous advantage of face validity in administrative or supervisory judgment tests suggests their use, if not in lieu of, certainly in conjunction with tests of vocabulary or general intelligence.

VIII. READING ABILITY

Since supervisors are frequently confronted with the necessity of reading policies, procedures, technical orders and the like, it seems reasonable to assume that the better supervisor may be a better reader. Colby and Tiffin reported testing 518 supervisors in seven Indiana manufacturing plants with the Nelson Silent Reading Test and found that they showed a mean grade level of reading ability slightly above the tenth grade.³⁷ While this does not establish validity for the test, it does suggest that a cut-off score weeding out the poorest readers might yield better supervisors, though a relatively high relationship with intelligence would seem to underlie this test.

IX. RATINGS AS PREDICTORS. SOCIOMETRIC AND SOCIODRAMATIC METHODS

Howell³⁸ postulated that " . . . group interrelationships might be measured and the relative positions of individuals within a group ascertained objectively by means of the sociometric test, which is basically a measure of the attractions and repulsions as between the various indi-

³⁷ Archie N. Colby and Joseph Tiffin, "The Reading Ability of Industrial Supervisors," Personnel, 27:147-153, September, 1950.

³⁸ Charles E. Howell, "Measurement of Leadership," Sociometry, 5:163-168, May, 1942.

viduals constituting the group."³⁹ There were two parts to Howell's study. The first was based on the Zeleny Group Membership Record in which each individual was given a list of names of the persons in the group and asked to indicate whether he liked, disliked or was indifferent to such activities as eating, working or playing with that individual. Each individual would get one point for each rating of like and would lose one point for each rating of dislike. The leadership score thus obtained correlated positively but low with an intelligence test ($r = .081$) and $.39$ with grade point school average. The second part of the study involved selection of twenty-five known leaders from 1100 students plus twenty-nine random students. These groups were rated by all 1100 students. The mean social status score for the leaders was 47.92 as compared with 28.99 for the random group. The correlation of this score with scholarship was $.42$; with intelligence $.69$. This study suggests the possibility of a group designating its own leader on the basis of his display of qualities associated with leadership while still in a non-leader situation.

³⁹Ibid., p. 163.

Eaton⁴⁰ asserted that tests and ratings, ". . . no matter how carefully constructed and administered, may never reach a high degree of relationship on a fine numerical scale. . . . Rough distinctions are perhaps more appropriate and valid for the evaluation of complex social personality 'traits' such as leadership."⁴¹ After reviewing tests applied to OSS and other intelligence groups, Eaton concludes that though the data are somewhat sketchy, they do tend to confirm that sociodramatic tests combined with sociometric ratings could be developed to separate upper and lower quartiles of persons most likely to succeed or fail as leaders in a specified group.

While the theory that leadership is specific to the situation or group to which it is applied sounds logical, Bell and French found by varying group membership that this accounted for a relatively small proportion of the variance in leadership status.⁴²

⁴⁰Joseph W. Eaton, "Experiments in Testing for Leadership," American Journal of Sociology, 52:523-535, May, 1947.

⁴¹Ibid., p. 534.

⁴²Graham B. Bell and Robert L. French, "Consistency of Individual Leadership Position in Small Groups of Varying Membership," Journal of Abnormal and Social Psychology, 45: 764-767, October, 1950.

As predictors, sociometric techniques or sociodramatic methods have not demonstrated validity sufficient to suggest their unlimited use. Something may be said for their use as criteria, however, against which to validate tests since the search for better methods of evaluating leadership success is most demanding.

X. RATINGS AS PREDICTORS: ADDITIONAL METHODS

In attempting to obtain rankings of a selected group of cadets at the U. S. Military Academy at West Point on leadership, Page found that, as determined by annual ratings of the senior class and the commissioned officer in charge of the companies, leadership ratings bore a closer relationship to bearing and appearance than to other subjects.⁴³

In developing a selection system for promotion of Sergeants to Captain of the State Highway Patrol, four techniques were applied by Browning, Hammond and Fenger; first, a personality assessment from a group interview with sociometric diagrams weighted 30% of the whole; second, a written knowledge test weighted 20%; third, an ex-

⁴³David P. Page, "Measurement and Prediction of Leadership," American Journal of Sociology, 41:31-43, July, 1935.

perience rating weighted 30%; and fourth, the organization merit rating weighted 20%.⁴⁴ While this study is more a case report in approach, it is worth noting that the selections based on this method had a validity of .84 when correlated with ratings made by former associates.

Williams and Leavitt, in an intensive study of the military corps found that ratings by colleagues were more valid than any other method for predicting leadership.⁴⁵

While the interview holds some promise as a selection device, Rundquist feels that its susceptibility to bias, the difficulty of finding a consistent frame of reference for expressing results and its time-consuming nature militate against its widespread use for evaluating individuals.⁴⁶

⁴⁴Rufus C. Browning, Kenneth R. Hammond and Frederic T. Fenger, "Self-Selection of Personnel," Public Personnel Review, 12:9-12, January, 1951.

⁴⁵Stanley B. Williams and Harold J. Leavitt, "Group Opinion as a Prediction of Military Leadership," Journal of Consulting Psychology, 11:283-291, November-December, 1947.

⁴⁶Edward Rundquist, Chapter IV, Section 29, "Personality Tests and Predictions," in Douglas Fryer and Edwin R. Henry, Editors, Handbook of Applied Psychology (New York: Rinehart and Company, 1950), p. 183.

XI. PERSONALITY MEASURES

In attempting to apply the Bernreuter Personality Measures to selection of supervisors, Hanawalt and Richardson found that most difference between supervisors and non-supervisors occurred in the Adjustment Scales, especially neurotic tendency, introversion-extroversion and self-confidence, and in Dominance, although this latter difference was not reliable.⁴⁷ The authors felt on the basis of their study that a test of leadership could be constructed with a key developed on items reflecting a significant difference between supervisors and non-supervisors.

Sparks also reports application of the Bernreuter Personality Measures in selection of supervisors.⁴⁸ Of 492 foremen in an oil refinery ranked by from four to twelve senior supervisors, 241 were uniformly rated, and of these, 191 took the Bernreuter Inventory. These were 90 first-line, 71 second-line and 30 third-line distributed in ratings as fol-

⁴⁷Nelson G. Hanawalt and Helen M. Richardson, "Leadership as Related to the Bernreuter Personality Measures: IV. An Item Analysis of Responses of Adult Leaders and Non-Leaders," Journal of Psychology, 28:397-411, October, 1944.

⁴⁸Charles P. Sparks, "Limitations of the Bernreuter Personality Inventory," Journal of Applied Psychology, 35: 403-406, December, 1951.

lows: 65 best, 66 worst and 60 average. Of the total number of items in the test, only thirteen items specifically keyed differentiated significantly but these were considered too few to be a reliable selection measure. When Flanagan's self-confidence and sociability keys and Richardson's scales were applied, they also failed to correlate with the criterion.

The Rorschach test has been applied in selection of officer candidates by Jensen and Rotter.⁴⁹ The sample consisted of fifty-six actually excellent officers selected by unit commanders at Armored Officers' Candidates School plus 257 officer candidates. Using a multiple-choice form of the Rorschach Test designed by Harrower and Erickson, it was found that the mean scores were practically the same for the excellent officers as for the officer candidates (3.4 as opposed to 3.2). No significant differences were found when the scores on the Rorschach were dichotomized (0-3 versus 4-10) and cross-tabulated with such tests as a Health Inventory, Psychasthenic Inventory or Group Level of Aspiration Test, purportedly measuring aggressiveness,

⁴⁹ M.B. Jensen and J.B. Rotter, "The Validity of the Multiple-Choice Rorschach Test," Psychological Bulletin, 42:182-185, March, 1945.

ambition, cautiousness and emotional stability.

Sinaiko reported utilization of another projective test, the Rozenweig Picture Frustration Test to select department-store section managers.⁵⁰ With a sample of 53, and a criterion of job efficiency measures from review of personnel data, Sinaiko obtained statistically significant negative relationships with the extrapunitive and need-resistance keys and positive relationships with the intropunitive and ego-defensive keys. By combining the scores, it was found possible to reject eleven of fifteen less successful managers and screen in ten of fifteen more successful managers.

The Humm-Wadsworth Temperament Scale was used by its authors who concluded that "On the basis of admittedly incomplete evidence, it would seem that a profile for executives would show the balance displayed by the average of the general population, with higher than average, but not extreme, tendencies in the elements designated as normal and manic."⁵¹

⁵⁰H. Wallace Sinaiko, "The Rosenzweig Picture-Frustration Study in the Selection of Department Store Section Managers," Journal of Applied Psychology, 33:36-42, February, 1949.

⁵¹D.G. Humm and G.W. Wadsworth, "The Humm-Wadsworth Temperament Scale," Personnel Journal, 12:314-323, April, 1934.

Freeman discussed a Pilot Behavior Blank used in the Army Air Forces to yield data on types of leadership based on Kurt Lewin's classifications of leadership: Laissez-faire (giving help only when requested to do so); authoritarian (dictatorial, domineering); and democratic (participating in a group on a peer basis).⁵² While the items were not validated, they suggest possibilities for future research.

The Office of Strategic Services in their selection of agents used an "organismic" approach, attempting to evaluate each personality as a whole.⁵³ The methods employed were sociodramatic and psychodramatic, including stress and post-stress interviews, interrogation tests, written tests and others. Attempts were made to measure motivation for the assignment, energy and initiative, effective intelligence, emotional stability, social relationships, leadership, security, physical ability, ability to observe and report, and propaganda skills. While it was admittedly difficult to determine validity and reliability, the following validities were obtained: selection against overseas staff appraisal, .37; with rating by theatre commander, .23;

⁵² Frank S. Freeman, Theory and Practice of Psychological Testing, (New York: Henry Holt and Company, 1949).

⁵³ Office of Strategic Services, The Assessment of Men (New York: Rinehart and Company, 1948), pp. 38-229.

with reassignment area appraisal, .08; and with returnee area appraisal, .19.

Personality measures of many types have been employed, mostly with negative or uncertain value. The need for better assessment of personality components is so great, however, that all research in the area of supervisory or executive selection would appear warranted. Rundquist voices the opinion that "A view widely held is that for supervisory positions personality characteristics are more important than skill or technical know-how."⁵⁴

XII. GENERAL STUDIES

File developed a series of items including supervisory judgment and attitude questions which he claimed were applicable to all industrial concerns in which he stressed supervisor-worker relationships as the key determiners of good or poor supervision.⁵⁵ These tests, the author

⁵⁴Edward Rundquist, Chapter IV, Section 29, "Personality Tests and Predictions," in Douglas H. Fryer and Edwin R. Henry, Editors, Handbook of Applied Psychology, (New York: Rinehart and Company, 1950) p. 182.

⁵⁵Quentin W. File, "The Measurement of Supervisory Qualities in Industry," Journal of Applied Psychology, 29 : 323-328, October, 1945.

claimed, could be used to predict the outcome of supervisory training programs. Later, File and Remmers applied the test to forty-six successful supervisors and fourteen non-supervisors in industry.⁵⁶ They found that the test differentiated successfully between these groups:

	<u>Number of Successful Supervisors</u>	<u>Number of Non- Supervisors</u>
Above 50th percentile	80	15
Below 50th percentile	20	85
Mean percentile score	75	23

Sartain administered File and Remmers' "How Supervise?" (Experimental Edition, Form A) together with the Otis Self-Administering Test of Mental Ability (Higher), Tiffin and Lawshe Adaptability Test (Form A), Revised Minnesota Paper Form Board, Bennett Test of Mechanical Comprehension (Form AA), Bernreuter Personality Inventory and Kuder Preference Record to forty supervisors in an aircraft plant.⁵⁷ Two ratings reduced to standard scores were the criteria used in the study. Correlations obtained were all below the acceptable level of significance, the highest being .18.

⁵⁶ Quentin W. File and H. H. Remmers, "Studies in Supervisory Evaluation," Journal of Applied Psychology, 30:421-425, October, 1946.

⁵⁷ A.Q. Sartain, "Relation between Scores on Certain Standard Tests and Supervisory Success in an Aircraft Factory," Journal of Applied Psychology, 30:328-332, October, 1946.

Jones and Smith worked with a general inventory of problematic questions, personal history questions and personality questions.⁵⁸ The sample consisted of seventy-two supervisors: twenty average in ability, twenty-two below average, and thirty above average. The point biserial correlation coefficients against this criterion of success were as follows:

Problematic questions	.43
Personality questions	.41
Personal History questions	.16
Total Battery	.46

Goode surveyed the literature on selection of leaders and found the following to be associated with successful leadership:⁵⁹

- (1) Mental ability
- (2) Breadth of interests and aptitudes
- (3) Language facility
- (4) Maturity
- (5) Motivation in terms of liking the work, ambition and perseverance
- (6) Social orientation in terms of realizing the need for cooperative effort
- (7) Reliance on administrative skills more than technical skills associated with his work

⁵⁸Omer R. Jones and Karl U. Smith, "Measurement of Supervisory Ability," Journal of Applied Psychology, 35: 146-150, June, 1951.

⁵⁹Cecil E. Goode, "Significant Research on Leadership," Personnel, 27:342-350, March, 1951.

It may be of interest at this point to review the foregoing survey of the literature dealing with supervisory selection to ascertain which psychological functions of the worker have shown promise in earlier studies as measurably related to supervisory success and which testing techniques have served validly to measure those qualities.

Intelligence has been demonstrated to be related to success when measured by the Otis Self-Administering Examination, Higher; verbal ability as measured by vocabulary and word series tests; reading ability as measured by the Nelson Silent Reading Test; interest as measured by a specially constructed key on the Strong Interest Inventory; supervisory judgment as measured by File and Remmers' How Supervise?, Jones' and Smith's problematic questions, and Mandell's administrative judgment test; supervisory attitudes as measured by File and Remmers' How Supervise? and Mandell's supervisory attitudes test; background characteristics as measured by a biographical information blank and Jones' and Smith's personal history questions; and personality as measured by G.W. and F.H. Allport's A S Reaction Study, the Rosenzweig Pictures and Jones' and

Smith's personality questions. As will be seen later, measures of ability, interest, supervisory attitudes, background characteristics and personality were included in the experimental battery as a means of taking advantage to the greatest extent possible of previous research in this field.

XIII. SUPERVISION DEFINED

As was noted earlier, few attempts have been made in most studies, experimental or otherwise, to define or delimit the group on which research was being performed. The terms "executive," "administrator," "foreman," "boss," "leader" and "supervisor" have, among others, been used interchangeably. Smith perceived the problem aptly when he remarked,

Not only are the group and individual factors that may affect the nature and effectiveness of leadership numerous and complexly interrelated, but as leadership studies have taken an increasingly prominent place in the literature, it has become evident that there is little uniformity in the operational meaning given to the central concept.⁶⁰

⁶⁰M. Brewster Smith, "Social Psychology and Group Processes" in Calvin P. Stone and Donald W. Taylor, Editors, Annual Review of Psychology, (Annual Reviews, Inc.: Stanford, California, 1952) p. 184.

Whether or not this has been partially responsible for the confusion and conflict in results of research studies conducted can only be a matter of conjecture. And yet it is a fundamental principle of test construction that the job for which the test is being developed be carefully analyzed and described. Selection of an experimental test battery is likewise based to a large extent on the philosophy of supervision within the organization, the duties and responsibilities of the supervisor, the leader-follower relationship and to some extent, situational components and organizational structure.

Social interaction and supervision. A landmark in the field of personnel research is the study accomplished at Western Electric from which a new concept of leadership emerged, indicating that the function of " . . . supervisors and managers was to listen to, and become better acquainted with, the sentiments of their employees and with the nature of that social structure, or system of sentiments, called 'the company.'⁶¹" . . . the social

⁶¹
F.J. Roethlisberger, Management and Morale (Harvard University Press: Cambridge, Massachusetts, 1946), p. 43.

structure of any particular company determines the kind of collaboration, the kind of people who will stay in the company, and the kind of people who will reach the top."⁶² Thus it was recognized that (1) workers were human beings, not commodities, and that recognition of them as human beings was a requirement of effective management, and (2) the organization was a social structure which itself could fashion the leader-follower relationship. Going on to determine the qualities needed for effective executive performance, Roethlisberger concluded "That a good portion of the executive's environment is verbal seems hardly open to question. . . . On the one hand, he has to become skillful in using words that will appeal to the listener's sentiments. . . . On the other hand, the executive has to be able to interpret skillfully what people say, for insofar as the work involves the interactions of human beings, his data come from what he hears as well as from what he sees and does."⁶³ Furthermore, Roethlisberger postulated a need for skill in diagnosing human situations. This analysis suggests strongly the use of pre-

⁶²Ibid., p. 45.

⁶³Ibid., p. 88.

dictors in supervisory selection studies to determine verbal-judgment ability and ability to understand workers as human beings and their problems as human problems.

Another approach to the description of leadership in social terms was made by Link who considered it an aspect of social effectiveness.⁶⁴ "Leaders are not merely born, they are persons who have acquired social effectiveness to an unusual degree. The popular belief that only a few are born to be leaders, the great majority must be followers (italics in the original) is a fallacy."⁶⁵ Link suggests that a successful and happy marriage, ability to get and hold a job and achieve promotion are elements of social effectiveness. Habits found in an individual's background which are related to social effectiveness and leadership include membership in organized and competitive groups, ability to be a good follower, practice in social skills, economic independence, bodily activity and dealing with the opposite sex.

⁶⁴ Henry C. Link, "Social Effectiveness and Leadership" in Douglas Fryer and Edwin R. Henry, Editors, Handbook of Applied Psychology (New York: Rinehart and Company, 1950), pp. 3-10.

⁶⁵ Ibid., p. 3.

The need for effectiveness in dealing with people is supported also by findings of Rupe who factor analyzed items in a rating made by subordinates on the Purdue Rating Scale for Administrators and Executives composed of 36 items in 10 logical groups.⁶⁶ Two factors emerged. The one which accounted for two-thirds of the variance dealt with behavior toward subordinates.

Theoretical analysis of the leadership function advanced by Brown substantiates the previous reports and provides a logical background for selecting factors to emphasize in developing leadership measures.⁶⁷ He specified five laws:

- (1) The leader must represent a region of high potential in the social field.
- (2) The leader must realize the existing field structure.
- (3) The really successful leader realizes the long-time trends in field structure.
- (4) Leadership increases in potency at the cost of decrease in freedom of leadership.
- (5) The successful leader must have membership character in the group he is attempting to lead.

⁶⁶J.C. Rupe, "When Workers Rate the Boss," Personnel Psychology, 4:271-289, Autumn, 1951.

⁶⁷J.F. Brown, Psychology and the Social Order (New York: Rinehart and Company, 1936), pp. 342-346.

This analysis forms the basis of what has apparently become the more generally applied democratic or equalitarian principles of management. It also suggests the need for abilities in perceiving the social structure in a working organization, willingness to subordinate one's personal demands to the group and predicting trends in organizational goals from information presently on hand.

Ohio State University Leadership Studies. The organizational emphasis was further stimulated by the Ohio State University studies in leadership. These studies took diverse forms, but primarily used modified job analysis and sociometric techniques to determine work patterns within organizations and to gain understanding of leadership as an organizational phenomenon. One study reported by Browne involved development of scales to measure responsibility assumed, authority exercised and authority delegated (R, A and D) by leaders in an organization.⁶⁸ Though only a small number of cases was used, it was possible to gain insight into the perceptions of executives regarding their manner of exercising these three aspects of their positions. One finding was that executives estimated the

⁶⁸C.G. Browne, "Study of Executive Leadership in Business: I. The R.A.D. Scales," Journal of Applied Psychology, 33:521-526, December, 1949.

authority they delegated to be less than either their own responsibility or authority. In another study, Browne developed a goal achievement index based on discrepancies between executives' ratings of the importance of various organizational goals and their ratings of the extent to which they believe the organization is achieving those goals.⁶⁹ The discrepancies are then expected to represent a lack of adequate communication and need for corrective action.

Shartle reported studies dealing with executives in colleges, industry and the military, concerned with dimensions of leader behavior, staff or group behavior and the broader cultural and economic environment surrounding the leader and group.⁷⁰ In one study, nine dimensions of leader behavior were set up a priori including initiation of ideas or practices, membership with the group, representation of group interests, integration of individuals with the group, organization, domination, communication, recognition of group members and production. Ten to twenty items were developed for each dimension and these

⁶⁹C.G. Browne, "Study of Executive Leadership in Business: III. Goal and Achievement Index," Journal of Applied Psychology, 34:82-87, April, 1950.

⁷⁰Carroll L. Shartle, "Leader Behavior in Jobs," Occupations, 30:164-166, December, 1951.

were given to 357 persons, of whom 152 were themselves leaders. It was possible to factor analyze the data and thereby identify three major factors: (1) behavior increasing the leaders' acceptability to the group, (2) objective attainment and (3) group interaction facilitation.

Hemphill, also engaged in the Ohio State University Leadership Studies, postulated that a definition of leadership would have to include characteristics of both the individual and the group.⁷¹ He compiled a list of fifteen fundamental, descriptive terms of group characteristics: cohesiveness, homogeneity, flexibility, permeability, polarization, autonomy, potency of involvement of members, stability, intimacy, degree of control over behavior of members, hedonic tone, association with members, participation of members, dependency on leadership, positions of members and size of group. Based on the responses of 500 persons to a questionnaire which asked them to describe in effect their group membership in terms of the 15 dimen-

⁷¹ John K. Hemphill, Situational Factors in Leadership, (Bureau of Educational Research Monograph Number 32: Columbus, Ohio, 1949), pp. 1-136.

sions Hemphill concluded that leaders of large groups behave differently from leaders of small groups. The reliability of the ratings was found to vary from .53 to .95 for the various dimensions. Although situational factors seemed to account for a large part of leadership success, five characteristics of leadership seemed to hold for all types of situations: (1) ability of the leader to advance the purposes of the group, (2) his administrative competence, (3) his pace setting and motivating activity, (4) his contribution to members' feeling of security about their place in the group and (5) his freedom from activities serving only his own interests.

The situational aspects of leadership were also stressed by Stogdill.⁷² Based on an exhaustive review of the literature, Stogdill concluded that the qualities, characteristics and skills needed by a leader are fashioned to a large extent by the demands of the particular situation in which he will perform as a leader.

Scott, working with Shartle, Browne and Hemphill concluded that " . . . the morale, if not the effective-

⁷²Ralph M. Stogdill, "Personal Factors Associated with Leadership: A Survey of the Literature," Journal of Psychology, 25:35-71, January, 1948.

ness of an organization may be dependent upon each member having a clear conception of his responsibilities and relations to his fellow members."⁷³ Enlisted naval personnel were asked to indicate the way they perceived organizational relationships and these were compared with the actual or official relationships. He found that perceptual errors tended to be lowest in those units in which the leader is high in authority, level and rank. Errors also tend to be low in units supervised by leaders who devote considerable time to consulting with associates and inspections of the organization. Leaders of low error units tend to work predominantly with peers, rather than with subordinates.

While the Ohio State University studies hold great promise as a basis for learning more about leadership as it functions in organizations, conclusions concerning what makes for effective leadership are as yet highly tentative. In the main, the studies are descriptive, hoping to give insight into the organization-leader relationship which may later, after refinement of methodology, perhaps provide the means for differentiating the potentially success-

⁷³Ellis L. Scott, Perceptions of Organization and Leadership Behavior (Ohio State University Research Foundation: Columbus, Ohio, 1952), pp. 1-109.

ful from the potentially unsuccessful leader. The practical difficulty of observing and describing each unit from which the sample would be drawn for this research as demanded by the situational approach, militated against giving it serious consideration. In accomplishing this research, therefore, the author was cognizant of the value of these studies but did not make direct use of them.

Furthermore, although situational aspects of leadership may represent important considerations for selection, it may be noted that Hemphill himself found certain characteristics to pervade all situations. This hypothesis is substantiated by Carter whose study on leadership and group behavior revealed that while group behavior differs with relation to leadership depending on the kind of task involved, a general leader ability does appear in all of the specific tasks.⁷⁴

Supervision and organizational effectiveness. Another attack on the problem of defining leadership, in this case in terms of group effectiveness, is being made

⁷⁴ Lanour Carter, W. Haythorn, E. Meirowitz and J. Lanzetta, "The Relation of Categorizations and Ratings in the Observation of Group Behavior," Human Relations, 4:239-254, Number 3, 1951.

by the University of Michigan Survey Research Center, where descriptive comparisons are being made of the supervisory tactics employed by those leading high or low-producing groups. Likert and Katz reported that high-producing sections were characterized by supervisors who are more employee-centered than work-centered, who delegate authority and give workers a sense of security in knowing where they stand with the company.⁷⁵ Higher morale units described their supervisors as creating a feeling of participation, as praising more than criticizing, as having interest in the worker, and as setting reasonable goals.

The criterion of organizational effectiveness was also employed by Comrey, Pfiffner and Beem, who asked 413 persons at 6 levels in 18 U. S. Forests in California in a questionnaire to describe the top administrator.⁷⁶

⁷⁵Likert and Daniel Katz, "Supervisory Practices and Organizational Structures as They Affect Employee Productivity and Morale," in Schuyler D. Hoslett, Editor, Human Factors in Management (Harper and Brothers; New York, 1951), p. 327.

⁷⁶A.L. Comrey, J.M. Pfiffner and H.P. Beem, "Factors Influencing Organizational Effectiveness: I. The U.S. Forest Survey," Personnel Psychology, 5:307-328, Winter, 1952.

The Forests were ranked independently in terms of effectiveness by a consensus of qualified personnel in the Regional Office. Forest supervisors in the more highly rated forests were found to be:

- "(1) more democratic with their top assistants, allowing them greater participation in running the organization.
- (2) more likely to interact socially with their top subordinates.
- (3) more likely to share information with their top subordinates.
- (4) more sympathetic in dealing with their top subordinates and their personal problems.
- (5) less critical of top subordinates and their work.
- (6) more critical of certain higher administrative policies.
- (7) more willing and able to help top subordinates in their work.
- (8) lower on longevity factors."

More highly rated forests were also found to have less dissension among their employees and to be run in an efficient but not autocratic manner.

The "employee orientation" of supervisors, suggested as related to group productivity, implies the value of a feeling for workers as human beings, an aspect of super-

visory performance which the supervisory attitudes and supervisory judgment tests are designed to measure, and which it will be seen, are included in the experimental test battery.

Critical incidents in leadership performance. Another approach to defining the leader's job was taken by the American Institute for Research under the aegis of John C. Flanagan. Flanagan and associates developed a concept of critical incidents or critical requirements in a leader's job which were judged to be "... those ways of performing which had 'made the difference' between being judged an effective officer or an ineffective officer by others in the organization."⁷⁷ From 3,000 such descriptive incidents collected by interviewing 640 officers, the final number was reduced to 77 by classification methods which related to performance of crucial functions. When used as a basis for developing an evaluation form for Air Force officers, the final product covered the following broad areas of proficiency: handling administrative details, supervising personnel, planning and directing action, skill in the military occupa-

⁷⁷Harley O. Preston, The Development of a Procedure for Evaluating Officers in the U.S. Air Force, (American Institute for Research: Pittsburgh, Pennsylvania, 1948) p. 68.

tional specialty, organizational responsibility and personal responsibility.⁷⁸

The technique of developing critical requirements for jobs, it can be seen, is a laborious one, not particularly practical to consider as a basis for test development in an applied setting. Furthermore, the value which underlay the technique seemed more promising as a means of criterion development and refinement than for experimental test selection, and since rating methods to be used appeared to assure reliable ratings, there seemed little advantage to be gained by pursuing this avenue of approach.

Supervision and the authoritarian personality. Still another attempt at describing leadership behavior was made by Sanford who postulated that " . . . leadership is a relationship between leader and follower, . . . that varies with (a) the behavior of the leader, (b) the predispositions and expectancies of the follower and (c) the supra-individual characteristics of the social situation in which leadership occurs."⁷⁹ Leaders must meet the needs

⁷⁸John C. Flanagan, "Job Requirements" in Wayne Dennis et al., Current Trends in Industrial Psychology (University of Pittsburgh Press: Pittsburgh, Pa., 1946), pp. 32-54.

⁷⁹Fillmore H. Sanford, Authoritarianism and Leadership, (Stephenson Brothers: Philadelphia, Pa., 1950) p. 3.

of followers and change with changing needs. Sanford's purpose was one of measuring the quality of authoritarianism and to examine the relationship to leadership orientation among followers. In measuring this quality, Sanford used a scale based on the work of Adorno et al. which described the authoritarian syndrome in terms of conventionalism, submission to authority, superstition and stereotypy, power and "toughness," destructiveness and cynicism, and the like.⁸⁰ Basically, the scale represented a measure of individual adjustment to authority. After using the scale in experimental studies, Sanford found that authoritarian followers appear to be more aware of the leader as a person, rather than as one who has a social function to perform; are neither interested in nor desire warmth and responsiveness to people in the leader; prefer strongly directive leadership; are not concerned with the welfare of their fellow followers; and demonstrate a concern for good solid character in the leader.

This approach holds promise as another basic indicator or way of describing the leadership situation. It does not claim that success is assured either the authoritarian or democratic leader, so that as a basis for assign-

80

T.W. Adorno, Else Frenkel-Brunswik, Daniel J. Levenson and R. Nevitt Sanford, The Authoritarian Personality, (Harper and Brothers: New York, 1950), pp. 262.

ing criterion ratings, it does not yet have value. Furthermore, the scale by which one determines this variable in any individual does not appear to be suited to selection work because of the rather obvious nature of the questions and the ease with which a desired answer may be predicted. This approach, therefore, did not appear ready for use in the present research.

Job analysis of supervisory positions. The job analysis approach to describing the supervisory job deserves mention since it serves so well the cause of experimental test selection in other occupational areas.

Moore described the functions of the supervisor as including representation of management to the worker and representation of the worker to management.⁸¹ He must aid in training his men. He must be on the alert for the occasional trouble-maker or malcontent who may get into his unit. And he must care for the controls and the paper work. Based on this analysis, Moore showed considerable insight into the measurements which could be best employed in selecting supervisors.

⁸¹Herbert Moore, "Supervision: I. Selection," Personnel Journal, 20:353-356, April, 1942.

Moore postulated that the modern foreman should be above the average in his level of mental ability. He continued that the foreman's ability to impart his information to others requires powers of analysis and ability to match imparted information with the capacity of the learner to absorb it, and he concluded that satisfactory personality characteristics were essential. He listed as the best potential predictors finally a general intelligence test, with different content or possibly different norms for different groups of supervisors; a personality test validated properly; and an interview.

Pfiffner assisted in defining supervisors by removing question of level from the other requirements.⁸² He defined supervisors as "all persons who are in formal control over others . . . irrespective of their high or low status in the hierarchy."⁸³

Mandell's analysis removed some of the confusion regarding the definition of a supervisor.⁸⁴ "By a supervisory position is meant one which involves responsibility

⁸² John M. Pfiffner, The Supervision of Personnel, (New York: Prentice-Hall Incorporated, 1951).

⁸³ Ibid., p. 6.

⁸⁴ Milton M. Mandell, "Testing for Administrative and Supervisory Positions," Journal of Educational and Psychological Measurement, 5:217-228, Autumn, 1945.

for the working conduct, and the quality and quantity of work produced by, one or more subordinates." . . .

"By an administrative position is meant one which involves extensive responsibilities for planning, organizing, directing, staffing, budgeting and coordinating the work of an organization or part of an organization."

. . . "Supervisory positions can generally be classified on the basis of three factors: (a) the number of employees supervised, (b) the nature of the work or the occupation supervised, and (c) the supervisor's level in

⁸⁵

the organization. From a survey of the literature, Mandell listed in addition to oral interviews and ratings in training classes as methods for selecting supervisors, the frequent use of paper-and-pencil tests. In his review, he noted interest inventories, personality inventories, tests of mental abilities, biographical information blanks, and such special tests as the Bennett Test of Mechanical Comprehension, Interpretation of Data, File and Remmers' "How Supervise?", Thurstone's Estimating Test and the Gottschaldt Figures Test. Mandell's

⁸⁵Ibid., pp. 217-218.

conclusion was that testing for supervisors should include a mental ability test and an interest inventory, and that more work was needed on all other forms.

A legal definition of supervision may be of interest to note. The Labor Management Relations Act of 1947 defines the term supervisor to mean " . . . any individual having authority in the interest of the employer, to hire, transfer, suspend, lay-off, recall, promote, discharge, assign, reward or discipline other employees, or responsibility to direct them, or to adjust their grievances, or effectively to recommend such action. . . ."⁸⁶

The situational aspects of leadership were stressed by Stogdill and Hemphill in separate papers.^{87, 88} Based on an exhaustive review of the literature, Stogdill concluded that the qualities, characteristics and skills needed by a leader are fashioned to a large extent by the

⁸⁶ Labor Management Relations Act of 1947, Section 2 (11).

⁸⁷ Ralph M. Stogdill, "Personal Factors Associated with Leadership: A Survey of the Literature," Journal of Psychology, 25:35-71, January, 1948.

⁸⁸ John K. Hemphill, "Situational Factors in Leadership," (Columbus, Ohio: Ohio State University, 1949) p. 136.

demands of the particular situation in which he will perform as a leader. Hemphill postulated that a definition of leadership would have to include the characteristics both of the individual and of the social situation.

Eisenberg defined supervisors as anyone responsible for directing the activities of others.⁸⁹ "First line supervisors are considered to be those individuals directly responsible for workers engaged in production, clerical operations, or transportation of materials." "'Coordinating supervisors' are individuals responsible for directing and reviewing the activities of first line supervisors."⁹⁰ Eisenberg next distributed a questionnaire to 374 individuals in 52 organizations, including military, industrial and commercial. Respondents were asked to check five qualities most essential to the first line supervisor. First line supervisors considered the following abilities most essential:

⁸⁹William J. Eisenberg, "Qualities Essential for Supervisors," Personnel Journal, 27:251-257, December, 1948.

⁹⁰Ibid., pp. 251-252.

- (1) Training and planning.
- (2) Assigning; delegating; developing teamwork; exercising authority and getting respect.
- (3) Improving job methods; knowledge of rules and regulations; skill in operations.
- (4) Maintaining records; evaluating results of operations.
- (5) Recognizing emotional disturbances; rating employees; planning future operations.
- (6) Selecting personnel; encouraging workers to grow; knowledge of the organization's standards of production; knowledge of related operations.
- (7) Skill in conducting group discussion.
- (8) Knowledge of organization's promotion policy, health and safety practices and plan for handling grievances.
- (9) Knowledge of organization's employment procedures, wage administration plan and special services.

Jones and Smith⁹¹ claim that a systematic approach to the analysis of supervisory and executive work suggest " . . . a four-fold division of performance as follows:

(1) knowledge of jobs supervised, (2) knowledge of technological controls applied to the jobs, (3) executive operations in planning and making decisions in the application of management operations to the work situation, and (4) leadership performance in dealing with the workers supervised.⁹²

⁹¹Omer R. Jones and Karl U. Smith, "Measurement of Supervisory Ability," Journal of Applied Psychology, 35: 146-150, June, 1951.

⁹²Ibid., p. 146.

Based on industrial and military experience, Freeman and Taylor described a valid and practical program for selecting leaders.⁹³

Speaking roughly, in leadership, personality factors are probably ten times as important as all aptitude and proficiency factors combined. . . . leaders are presumably distinguished from non-leaders by a peculiar relation of drive and adaptability factors, those elements which seem to contribute most to the effective mobilization and release of energy in interpersonal situations.⁹⁴

When the authors performed desk audits of business leaders, they observed three fundamental criteria of success: degree of acceptability, ability to delegate and decisive adaptability. The authors postulate that a " . . . logical, intellectual appreciation of what to do is not sufficient to get it done. A man may know (italics in the original) exactly how to act in a given situation, but his emotions and interests may lead him to a different procedure. . . . At our present stage of knowledge we would do well to include a good test of administrative judgment along with tests of non-verbal, non-social reasoning. . . .⁹⁵ The

⁹³G. L. Freeman and E.K. Taylor, How to Pick Leaders, (Funk and Wagnalls Company: New York, 1950), pp. 1-226.

⁹⁴Ibid., p. 18.

⁹⁵Ibid., p. 108.

authors went further to conclude that " . . . the outstanding success . . . can best be produced in the man who couples exceptional aptitude for verbal manipulations and administrative judgment with exceptional interest in persuading others to go along and in affecting group action."⁹⁶

Values derived from description of the supervisory job. A review of the literature dealing with attempts at understanding, analyzing and describing supervisory or leadership jobs has indicated the diversity of methods being used. There seems to be stated agreement that personality aspects of these positions is of great importance and implicit agreement of their relationship to success. The ability to understand sentiments of employees; recognition of a social structure; ability to diagnose human situations; effectiveness in social situations; ability to assume membership character in the group being led; acceptability within the group; ability to facilitate group interaction; ability to contribute to group members' feeling of security; possession of employee-centered attitude; ability to meet the needs of followers; and skill in human relations are all terms used to describe different aspects of the same phenomenon. The experimental test battery, it

⁹⁶Ibid., p. 115.

will be seen, has been designed to give considerable emphasis to prediction of personality characteristics through incorporation of supervisory problems, personal preference inventories and a biographical information blank. The heavy verbal aspects found in the job are measured by the supervisory judgment test which also covers to a large extent the judgmental abilities required for planning, decision-making, controlling, delegating, evaluating and coordinating. The emphasis placed on interest led to inclusion of an interest measure in the final battery.

CHAPTER III

THE CRITERION MEASURE

One of the most perplexing problems facing the personnel psychologist today is that of locating a satisfactory yardstick of job success against which to evaluate experimental predictors. Much of the difficulty stems from the inability to find objective indices which represent a worker's total satisfactoriness on the job. This is particularly true in the case of the supervisory job where objective measures are not likely to be available. Reliance must most often be placed, therefore, on some evaluation measure using a judgment of success. The limitations and criticisms of rating methods are well known, involving as they do the subjective element which may lead to ratings based on stereotypes, particular biases or prejudices, favoritism, and the "halo" effect.¹ Certain attempts at improving rating methods have been more or less successful, primarily involving the training of raters, assuring that they

¹William E. Mosher, J. Donald Kingsley and O. Glenn Stahl, Public Personnel Administration (New York: Harper and Brothers, 1950), p. 368.

are both willing and able to rate the subject, and asking for judgments which they are really in a position to make. One of the commonest weaknesses of rating methods, particularly the graphic rating scale which makes it less valuable for research purposes, is the bunching of ratees at the high end of the scale, a phenomenon probably due in large part to a supervisor's unwillingness to make negative evaluations of employees he is still associated with and the fear that such negative evaluations reflect on his own competence.

I. INGREDIENTS OF SUCCESS AS AN AIR FORCE SUPERVISOR

In the preceding chapter various approaches to description of the supervisory job were presented. Though possibly an eclectic approach to description provides the best picture of what a supervisor does and may suggest the potentially most predictive experimental battery, there still remains a gap between the "what" and "why" and the "how well." Yet an accurate evaluation of "how well" each supervisor in the sample is performing his supervisory duties is essential to development of any instrument for predicting that success.

Before one proceeds to develop methods for obtaining evaluations of supervisory effectiveness, however, it is necessary to specify for the given situation, what actually represents a successful supervisor. Insofar as the Air Force is concerned, a successful civilian supervisor is one who motivates his subordinates to reach organizational goals with highest levels of production and with maximum efficiency of manpower, money and materiel, retaining all the while a team spirit and the satisfaction of employees with their jobs.² This definition may be expanded to include such specifics as maintenance of effective relationships with peers and superiors, ability to recognize training needs, ability satisfactorily to assign work and evaluate performance, ability to dissipate problems of employees at his own level to the satisfaction of the employees, ability to develop economical work methods, ability to keep absenteeism and turnover to a minimum and interest in safety practices.

²From discussions with Mr. Jack E. Ehrmantraut, Chief, Career Planning and Training Division, Directorate of Civilian Personnel, Hq. USAF.

An official Air Force listing of major responsibilities and component tasks of a supervisor appears in the Air Force Advanced Supervision and Management Trainers' Course as follows:³

1. Understand the duties and responsibilities of his position.
2. Plan how best to accomplish the mission of his unit, determine as far as he can what is needed to accomplish it and make recommendations to supervisors.
3. Assign work and supervise others.
4. Evaluate and improve work methods, processes and procedures.
5. Improve his own knowledge, techniques and skills.
6. Determine the areas in which subordinates need training and provide it.
7. Evaluate employee performance and unit output in relation to abilities.
8. Assist employees to keep informed, to adjust their problems and to develop good discipline.
9. Help each employee to meet his obligation to the organization.
10. Carry out the policies, regulations and procedures of the organization.
11. Work with and coordinate his activity with that of other supervisors and officials.

³USAF School for Civilian Personnel Administration, Headquarters, USAF, Advanced Supervision and Management Trainers' Course, December 1949.

12. Recognize the need for assistance and call on supervisors and staff specialists for that assistance.

13. Keep superiors informed.

Within the Air Force, the ultimate objective is the maximum possible delegation of managerial responsibilities to the supervisor consistent with legislative requirements and good administrative practice. Supervisory training programs are developed around that concept and actually provide the instructional basis for acceptance of supervisory responsibilities. A program of planned assistance to supervisors is directed specifically at assisting supervisors to assume their full responsibility for personnel management. The result of this research study is expected to provide supervisors who will have the potential to absorb the training and profit from the assistance and thereby assume fullest possible managerial responsibility, at least in the personnel management area. The specifics of the training and assistance program in the aggregate represent the standard of supervisory success.

Superiors of the supervisors, having been exposed

to the training and assistance programs and having been indoctrinated with the Air Force concept of successful supervision, will utilize this standard, tempered by awareness of realistic situational limitations, in providing evaluations of their subordinates' success. Subordinates of the supervisors are provided a similar framework based on the features of supervision which they are in a position to evaluate, which mesh with the responsibilities and duties included a priori in the training and assistance programs. Against these criteria, the experimental predictors will be evaluated.

2. The diversity of functions represented by supervisors to be included in the sample and the general unavailability of objective measures of success of supervisors required the acceptance of subjective estimates of success for use as criteria in this research study. On the other hand, the following assurances were to be made that the criteria thus obtained provided the maximum of reliable information about the supervisors in the sample, yet also sufficient information about themselves to indicate its limitations.

II. PRECAUTIONS TO BE TAKEN IN OBTAINING RATINGS

Ratings would be obtained from both superiors and subordinates. Studies accomplished by Katz at the University of Michigan Survey Research Center suggested the bifurcated view of the effectiveness of a supervisor and it is a logical concept to follow.⁴ The supervisor to be successful must be both management- and employee-oriented. As a representative of management, he is constantly evaluated by his superiors in terms of his ability to meet management's goals. On the other hand, in his relationships with subordinates, his ability to conduct personnel management responsibilities can be more directly evaluated by those supervised. Freeman & Taylor suggested the desirability of evaluating leaders in terms of competence adjudged by superiors and subordinates.⁵ It was deemed desirable in this study, therefore, to obtain two sets of evaluations representing two possibly different aspects of a supervisor's success: the one from his superior, the other from his subordinates.

⁴Harold Guetzkow, Editor, Groups, Leadership and Men (Pittsburgh, Pa.: Carnegie Press, 1951), pp. 68-85.

⁵G.L.Freeman and E.K. Taylor, How to Pick Leaders (Funk and Wagnalls Company: New York, 1950), p. 40.

Ratings would, whenever possible, be collected from more than one individual, whether at the subordinate or superior level. This was necessary to enable a check to be made on the reliability of the criterion.

Raters must be willing and able to rate the supervisor concerned. Only if a superior or supervisor voluntarily provided the information requested would his ratings be included. This was necessary for two reasons: First, a rating provided under duress might be expected to be an unreliable one. But second, in the practical setting of a federal agency, it is not considered sound employee-management relations to invoke authority when the reason for doing so is purely experimental. The test of ability to rate was the subordinate's or superior's answer to the question, "Do you know the ratee well enough to evaluate his job performance?" An additional control was imposed by the requirement that a supervisor, to be included in the sample, must have been in the position of supervisor for at least six months, to assure that those evaluating him had been able to observe him for a period of time sufficient to make the desired evaluation.

Raters should be trained in advance of filling out the ratings. Advance training was provided all raters so that (1) all raters would be providing ratings under standardized conditions and with identical frames of reference; and (2) the research nature of this study could be explained along with the fact that the rating would not be used for any but experimental purposes, so critical evaluations, where appropriate, could be given under confidential circumstances; and (3) every means possible could be used to defeat the tendency to bunch ratings at the high end of the scale.

Collaboration in filling out ratings would be avoided. This would be accomplished by scheduling all sessions so that ratings would be prepared in a central place, and by having each rating session monitored to prevent exchange of information.

Superiors would be asked both to rate and rank their subordinate supervisors. By asking for both ratings and rankings from superiors, it could be demonstrated that though all their subordinates might be "good" some were better than others. Rankings accomplished first, then, might provide greater spread among the ratings made. Rank-

ing could not, of course, be accomplished by subordinates, since each subordinate would have only one supervisor whom he could rate.

The frame of reference for subordinates giving ratings would be standardized. Subordinates would be given a number of preliminary questions to answer about how well their supervisors handled specific aspects of their jobs before being asked to give an overall rating. This would assist in standardizing the configuration developed in the raters' minds in preparing them for the important job-success question.

Ratings would disregard technical competence and emphasize only supervisory ability. Every effort would be made to divorce technical competence from the rating and to disregard the ability of the supervisor to perform the purely technical function he chances to be supervising. This is important since the purpose of the study is to develop a battery of tests able to predict supervisory success without consideration for the activity supervised. That this effort was successful is reflected in the fact that after a repeat caution in this regard, nearly ten per cent of the raters revised their ratings who had based

the original mostly on technical competence.

Attempts would be made to detect raters with reading handicaps. Raters with reading handicaps would be located so that their ratings would be usable and as accurate as possible. Persons so detected would be given personal assistance in filling out their ratings. As might have been expected, reading difficulties were found with subordinates rather than with superiors, and the extent to which this created error variance is difficult to ascertain in view of the range of reading ability which exists between marginal ability at any given grade level and complete illiteracy.

Relevance to supervisory success would be built into the rating form. Referencing Thorndike, "The quality designated as relevance to the ultimate goal is the prime essential of a criterion measure. . . . That is, the systematic, non-error sources of variance in score on the criterion measure should arise from the same factors that make for ultimate success on the job, combined with the same weights."⁶ Though Thorndike recognized that this ultimate criterion is never achieved, all efforts made in this study were to approximate it.

⁶Robert L. Thorndike, Personnel Selection (New York: John Wiley and Sons, Inc., 1949), p. 125.

III. METHOD OF OBTAINING SUPERIORS' RATINGS

The mechanics. From personnel records, organization charts and knowledge of placement advisers and position classifiers in the civilian personnel office, a list was prepared of the officials known to be the superiors of those supervisors selected to be in the experimental sample. A second level of supervision was also listed. Having been prepared through prior publicity given to this study in installation newspapers, personnel on the list were expecting to be contacted. A schedule of group sessions was then drawn up and coordinated with those on the list. The purpose of the session was explained as having to do with obtaining information in connection with a research study aimed at improving Air Force methods of selecting supervisors. Raters verified their ability and willingness to rate their respective subordinate supervisors. Group rating sessions were designed to have up to eight raters present at a session. Small groups were arranged intentionally to permit as much individual attention for each rater as he felt was necessary. The detailed instructions were mimeographed and made available

to each rater. A copy of the instructions, entitled "How to Use the Employee Evaluation Blank" is attached as Appendix A. Briefly, it asks the rater to rank all subordinate supervisors in his organization, and then to rate them on a scale marked off with nine unnumbered blocks.

The ranking procedure. The method of ranking is somewhat unusual and merits description here. Dennis and Shartle refer to the technique as "the alternative ranking method."⁷ Raters are given 3 x 5 cards, each with the name of one of his subordinate supervisors, and asked to pair the employees. The first pair is to represent the very best and very worst supervisor of the group. The next pair represents the second best and second worst. The procedure continues until all members of the group have thus been ranked. This procedure permits the rater to recognize the sharpest possible differences in his mind between employees without forcing him to indicate how good is the best and how bad is the worst. The rating provides a check on the ranking, and in addition, specifies where on the scale each man stands, so

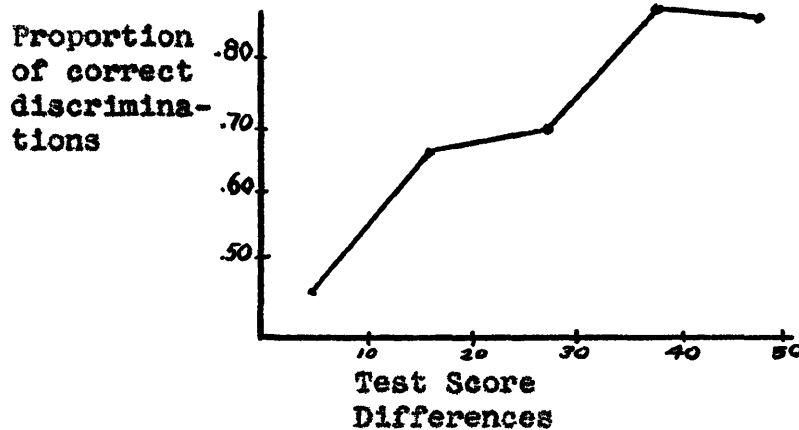
⁷Wayne Dennis, Carroll L. Shartle et al., Current Trends in Industrial Psychology (Pittsburgh, Pa.: University of Pittsburgh Press, 1949), p. 64.

that comparisons between groups are facilitated. An interesting excursion into the potentiality of the "alternative ranking method" was shown in the following:

Three test scores: an "unconventional" personality test, supervisory Judgment and Street-Gestalt - each converted into standard scores with a mean of 50 and standard deviation of 10 were available for 130 pairs of supervisors ranked on the above basis. The differences in these scores should theoretically be in favor of the higher ranking supervisor of each pair. The ability of the tests to make this differentiation is confirmation of validity ascertained by more complex methods. Table I on the following page demonstrates the ability of the test to predict wide differences in ranking and its ability to predict narrower differences somewhat less effectively.

TABLE I

ABILITY OF THREE TESTS TO PREDICT THE
CRITERION OF ALTERNATIVE RANKING



Cooperation of raters. Cooperation in securing superiors' ratings was excellent. Some difficulty existed in attempting to get superiors to rank and rate supervisory success apart from technical competence, but repeated emphasis on this point during orientation to the rating session, it is felt, largely eliminated this consideration from the final ratings submitted.

IV. METHOD OF OBTAINING SUBORDINATES' RATINGS

Mechanics. Subordinates likewise were acquainted through base publicity of the research study being conducted. Arrangements were made through supervisors in

the sample to permit their employees to assist in the study. Cooperation both of the supervisor being rated and of the subordinates doing the rating was excellent. The subordinates' ratings were collected in the form of an "attitude survey" which the base hoped would shed some light on the quality of supervision and suggest areas of supervisory methods in need of improvement. If a supervisor was responsible for ten or fewer employees, all were invited to participate in evaluating him. If he supervised more than ten employees the group was sampled to provide at least ten. Instructions were mimeographed and presented to all raters at each group session. The instructions were also read aloud. A copy of the "Instructions for Filling Out Questionnaire on Supervision" appears as Appendix A along with a copy of the questionnaire. Note that subordinates were asked not to identify themselves by name anywhere on the questionnaire. Questions 1 through 26 were used to educate subordinates to a standard configuration of thinking about supervisory effectiveness, based on the Air Force standard of good supervision and on those qualities subordinates would have an opportunity to observe. Actually, question 27 was the "payoff" in terms of the

evaluation measure used for the research; questions 1 to 26 served the additional purpose of providing "attitude" data for other uses. Question 27 is a nine-point graphic scale with a built-in attempt to keep raters away from the top end of the scale.

V. DISTRIBUTION OF RATINGS

The distribution of both superior and subordinate ratings reflects success in getting voluntarily a spread on the ratings, a feature basic to employment of the statistical technique of correlation. Tables II and III below disclose the theoretical frequency of ratings to be expected by chance (in terms of stanines - standard scores with a mean of 5 and standard deviation of 1.9) and the obtained distributions for the samples combined from both bases.

TABLE II

DISTRIBUTIONS OF RATINGS COMPLETED BY SUPERIORS AT BOTH INSTALLATIONS. (OBTAINED AND THEORETICAL)

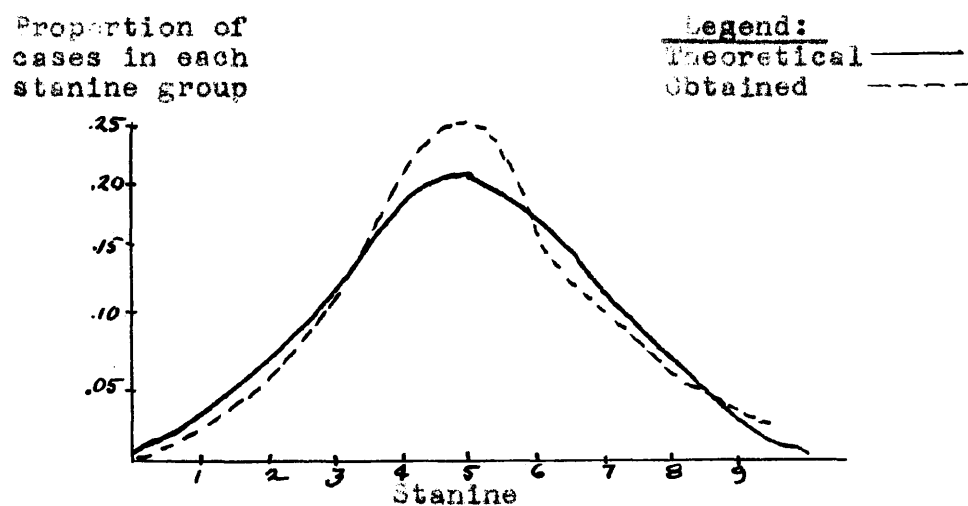
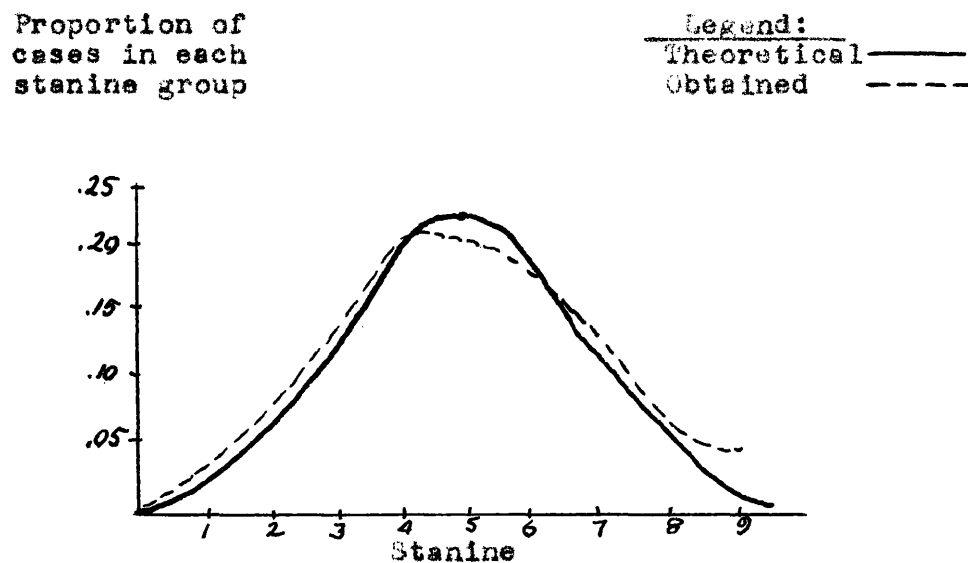


TABLE III

DISTRIBUTIONS OF RATINGS COMPLETED BY SUBORDINATES AT BOTH INSTALLATIONS. (OBTAINED AND THEORETICAL)



VI. RELIABILITY OF RATINGS

To check on the reliability of both superior and subordinate ratings, the following procedure was adopted. The superior ratings for each subject were split randomly into two groups and the mean rating computed for each half. This process was continued for each subject, providing him with two mean ratings based on a split of all ratings received. The same procedure was used with subordinate ratings. Table IV depicts the reliability coefficients for the superior ratings and for the subordinate ratings as well as their inter-relationships.

TABLE IV
RELIABILITY OF SUPERIOR AND SUBORDINATE RATINGS BY BASE

	Superior vs Superior	Subordinate vs Subordinate	Superior vs Subordinate
Tinker AFB	.372 N=230 *.542	.338 N=221 *.554	.254 N=229
Hill AFB	.304 N=151 *.554	.537 N=181 *.698	.163 N=189
Total			.211 N=418

* Spearman-Brown Prophecy Formula

$$r_{nn} = \frac{nr_r}{1 + (n - 1)r_r}, \text{ where}$$

r_{nn} is the correlation between n forms
of the test and n alternate forms.

r_r is the reliability of the test.

Garrett indicates that "the prophecy formula may be applied to ratings, estimates and other judgments."⁸

Thorndike points out that "It is more important that the reliability of a criterion measure be known (italics in the original) than that it be high (italics in the original).

⁸ Henry E. Garrett, Statistics in Psychology and Education, (New York: Longmans, Green and Company, 1938) p. 315.

This information is needed to establish the fact that it is not zero."⁹ Further, "High reliability in a criterion measure is convenient but not critically important."¹⁰ Low reliability in a criterion measure merely attenuates all its relationships with other measures.

. . . To compensate for these influences it is necessary to base validity correlations for unreliable criteria on substantially larger populations, so that each component statistic is determined with greater precision.¹¹ The reliability coefficients reported in Table IV are all significant at the one percent level of confidence, so that the criteria appear satisfactory for use in this study.

VII. COMBINING THE RATINGS

At this point in the study, practical problems of time and economy dictated the need, if at all possible, of selecting only one of the two criteria for use, either the superior or the subordinate. Rather than lose the information provided by each, however, the decision was made to take partial advantage of both by combining them.

⁹Robert L. Thorndike, PERSONNEL SELECTION (New York: John Wiley and Sons, Inc., 1949), p. 107.

¹⁰Ibid., p. 127.

Since the relationship between criteria was better than zero, it was considered appropriate to combine them.

Wherry confirms the acceptability of this procedure.¹¹

"If the various criteria are actually independent, as shown by zero or near zero intercorrelations, any composite function will be more or less meaningless. If, on the other hand, the criteria show uniform and moderate to high intercorrelations, some form of composite is indicated." Both criteria consequently were converted into stanine scores and the mean of the sum of subordinate and superior used as the final criterion. Stated another way, the final criterion was equivalent to the average superior rating plus the average subordinate rating divided by two.

This compromise would of course have least damaging effect where superiors and subordinates agreed on the performance of the supervisor rated. In other cases, the tendency would be to have ratings which differed regress toward the mean. If the purpose of the resulting instrument is to provide a means of screening out those with poorest potential in terms either of subordinate or superior ratings, or of screening in those with highest potential in

¹¹Robert J. Wherry, Chapter IV, Section 27, "Criterion and Validity," in Fryer and Henry, Editors, op. cit., p. 176.

both directions, the combined criterion will serve well. The error may exist where difference of opinion exists. In those cases, some would be screened in who may not deserve to be, and some would be screened out who did not deserve to be. Since management chooses to select personnel for supervisory positions who meet both criteria, the combining of criteria is a method of providing such personnel with a minimum of administrative difficulty.

Reliability of Combined Criterion. Taking half the subordinate ratings at random and half the superior ratings at random for each employee in the sample, and finding the mean of the resulting total; then doing the same for the other half, a correlation of the two halves should reflect the reliability of the rating thus combined. This procedure yields possibility of correlation in two ways: by combining superior and subordinate ratings for the first half and relating them to superior and subordinate ratings of the second half; or by taking superior of first half with subordinate of second half and relating them to subordinate of first half and superior of second half. Both were accomplished to permit obtaining the best estimate of reliability. Based on the first approach, relia-

bility corrected by Spearman-Brown Prophecy Formula was .55; accomplished the second way, it was .66. Applying the Fisher z transformation, the reliability of the combined criterion based on both approaches was found to be .60. A reliability of this order would be satisfactory for group prediction, which is an aspect of the basic purpose of this research study. Individual prediction would require higher reliability.

CHAPTER IV

THE SAMPLE

The selection of a human experimental sample in a large organization is a problem which often calls for compromises between theoretically ideal sampling procedures and practical administrative feasibility. Certainly if the assignment is one of developing tests for use anywhere in the Air Force, the sample chosen should be a random one of all Air Force Bases. Secondly, if the assignment is one to develop a test for all supervisors, the sample should include a random sample of supervisors within each selected activity. The problems existent in each of these steps are such that psychologists in similar studies have had to content themselves with approximations to the ideal.

For one, the cost of conducting a study at even one Air Force Base is high and an administrator can hardly be expected to multiply that cost by the addition of other installations merely for experimental purposes. These costs include time away from productive work of employees selected to take tests and persons called upon to provide evaluations of the success of that sample; time required in educating operating officials to the purpose and requirements of the study; and time of the

personnel office in selecting the sample, scheduling the persons to be tested and to provide ratings and administering the experimental battery. An additional cost would lie in the need for treating each installation separately for analytical purposes.

The second problem relates to the selection of a sample of supervisors within any one base. To keep the cost and administrative inconvenience to an installation to the very minimum, and to keep the time requirements down to those a personnel office can meet as well as those the research activity can meet, the smallest number of employees (in this case supervisors) must be contained in the sample. A further consideration lies in the important feature of obtaining a sample for whom at least two ratings from superiors will be available. When the criterion involves ranking of employees or paired comparisons, such as the one utilized in this study, it behooves the research investigator to keep the sample organizationally restricted. Thus it is most desirable to include in the sample all supervisors in a given division within which comparisons may be made between those supervisors by higher levels of management. With one supervisor taken from each division, comparisons become more difficult to obtain because of the unlikelihood that

higher levels of supervision will have direct access to or observation of supervisors in divisions outside their control.

1. TYPE OF SAMPLE

Deming defines two types of samples: probability and judgment.¹ Probability samples are those ". . . for which the sampling errors can be calculated, and for which the biases of selection, non-response, and estimation are virtually eliminated or contained within known limits." Judgment samples, on the other hand, are those "for which the biases and sampling errors cannot be calculated from the sample, but instead must be settled by judgment." Deming of course favors the probability sample and for many investigators, particularly of census or opinion polling varieties, it is not only desirable but feasible. While the desirability of selecting a sample according to a statistical plan is not to be questioned, the value of applying judgment when random selection procedures are not feasible must be recognized.

¹ William Edwards Deming, Some Theory of Sampling, (New York: John Wiley and Sons, Inc., 1950) p. 10.

II. SELECTION OF BASES

The resources of time, budget and administrative tolerance permitted the inclusion of two bases in this study with a sample of approximately five hundred supervisors. The bases selected were those representative of the Command employing the bulk of all civilians in the Department. The bases, Hill Air Force Base and Tinker Air Force Base, at Ogden, Utah and Oklahoma City, Oklahoma respectively perform missions similar to those of eight of the largest bases within that Command. Geographically they represent two distinct parts of the country. Table V below, however, indicates the similarities according to various factors in the sample drawn at Hill and Tinker Air Force Bases.

TABLE V
POPULATION STATISTICS

		Hill AFB	Tinker AFB	Total
AGE	Number	216	251	467
	Mean	37.0	40.7	39.0
	Standard Deviation	11.1	10.9	11.0
BLUE COLLAR GRADE	Number	143	163	306
	Mean	14.9	16.3	15.7
	Standard Deviation	4.0	3.6	3.8
PROFESSIONAL AND CLERICAL GRADE	Number	65	74	139
	Mean	5.9	6.0	6.0
	Standard Deviation	1.5	2.1	1.7
EDUCATIONAL LEVEL	Number	214	251	465
	Mean	11.8	11.5	11.6
	Standard Deviation	2.3	2.3	2.3

Within the installation, based on considerations discussed earlier, organizations were sampled and the universe of supervisors included from each organization.

III. SELECTION WITHIN BASES

In selecting organizations, an attempt was made to include professional, clerical and blue-collar supervisors in approximately the proportion they existed within the Air Force. Because of the sampling of organizations, however, the final numbers showed the blue-collar groups to be over-sampled, while the white-collar groups were under-sampled. Table VI below depicts the numbers finally incorporated in the sample by broad occupational groups and by base.

TABLE VI
OCCUPATIONAL DISTRIBUTION OF SUPERVISORS IN SAMPLE BY BASE

	Professional	Clerical	Wage-Board	Total
Hill AFB	37	30	150	217
Tinker AFB	22	52	178	252
Total	59	82	328	469

Though five hundred supervisors were initially planned for, absenteeism, incomplete data, turnover and inability to be spared accounted for the loss to 469. This loss has been assumed to be non-biasing.

Though five hundred supervisors were initially planned for, absenteeism, incomplete data, turnover and inability to be spared accounted for the loss to 469. This loss has been assumed to be non-biasing.

IV. CRITERIA FOR INCLUSION IN THE SAMPLE

It seemed that best results in sampling supervisors would be obtained by following leads suggested by Mandell which involve specifying the supervisors to be included in such terms as number of persons supervised, occupations supervised and the like.² The results would more likely be standard and objective and allow later specification for use of the resulting battery. This seemed more realistic and practical than attempting to analyze each alleged supervisory job in terms of a list of job duties. To be included in the sample, a supervisor must consequently have met the following criteria:

- (1) Must supervise persons engaged in work in which he is himself technically proficient.
- (2) Must supervise at least three of the kind of employees described in (1).
- (3) If a blue-collar supervisor, must not be above the grade of foreman; if clerical supervisor, not above the second line of supervision; if professional, not above the first line of supervision.
- (4) Must have been a supervisor for a least six months, so that his supervisory qualities could have been observed long enough for reliable rating purposes.

² Milton M. Mandell, "Testing for Administrative and Supervisory Positions," Journal of Educational and Psychological Measurement, 5:217-228, Autumn, 1945.

The job analysis approach in the strictest sense was not used in selecting the sample, the definition of supervisory duties and responsibility being subject to considerable difference of opinion. Instead, supervisors were selected on the basis of a perceptual distinction by both superiors and subordinates plus an operational distinction formed in the personnel office.

The conduct of personnel research is always difficult because it disrupts going administrative activities. The writer feels that this study nonetheless meets the criteria of size and appropriateness of sample, even if it does not meet all the criteria of a true probability sample. While the sample was chosen partly because of convenience, a conscious effort was built into the sample design to obtain representativeness. This approach may consequently be considered a compromise between Deming's probability sample and judgment sampling. The guiding principle in selecting the sample was that it be heterogeneous as to type of work supervised but homogeneous in terms of supervisory status. Thus all met the criteria described earlier in this chapter and at the same time represented a wide variety of occupations, including plumbing and steam-fitting, woodworking, engine spray painting, flight test mechanic, tabulating machine operator, statistical clerk, property and supply officer, training officer, and many others.

V. CHARACTERISTICS OF THE SAMPLE

Of the 469 supervisors for whom complete data were obtained, 428 or 91.2% were male; 41 or 8.8% were female. The average age was 39 years. The average educational level was 11.6 years, or close to a high school education. The average grade of the classified employees (professional and clerical combined) was GS-6; the average grade of the blue-collar employees was GS-15.7. Information concerning religion or race of supervisors was not requested because of federal practices regarding non-discrimination.

CHAPTER V

THE EXPERIMENTAL BATTERY

The following list of tests used in other studies, gleaned from a review of the literature plus unpublished studies being accomplished in other organizations suggests the wide variety of attempts made to isolate supervisory abilities and measure supervisory success:

1. Administrative judgment
2. Arithmetic reasoning
3. Attitude inventories
4. Blueprint reading
5. Clerical aptitude
6. Current events knowledge
7. Decision-making ability
8. Human relations understanding
9. Insight into others' personality
10. Interests
11. Interpretation of data
12. Mechanical ability
13. Mental ability
14. Numerical facility
15. Physical condition
16. Personal history data
17. Personality
18. Reading comprehension
19. Reading speed
20. Rules and Regulations knowledge
21. Social intelligence
22. Spatial visualization
23. Statistical estimation
24. Supervisory judgment
25. Trade knowledge (subject matter)
26. Verbal fluency
27. Vocabulary

Research reported in Chapter II suggested the value of certain aptitude and personality tests as predictors of success in the supervisory job. In Lawshe's terms, "Every supervisor has things to learn; every supervisor has some 'paper work' to do. Both of these facts suggest the importance of mental ability in the supervisor's job Similar to the selling field, instruments that measure attitudes, beliefs and interests seem most promising at present for selecting supervisors."¹ Previous studies accomplished under the writer's direction within the Air Force on junior administrators confirmed the significance of measurement of mental ability and "personality" characteristics.² Administration of the U. S. Employment Service's General Aptitude Test Battery yielded valid predictors only for verbal and spatial tests, both of which comprise the G or general intelligence factor in that battery. Evidence from all sources therefore, suggested the need for more intensive research into the personality sphere that was needed, or would be profitable perhaps, in the aptitude or knowledge spheres. A further limitation in choice of instruments was made by the administrative determination to keep the experimental battery to a four-hour time limit.

¹ Charles H. Lawshe, Jr., Principles of Personnel Testing, (New York: McGraw Hill Book Co., 1948) p. 171.

² Joseph G. Colmen, G. O. Fiedler and J. R. Blackburn, "Identification of Executive Talent within a Federal Department," paper read at national convention of American Psychological Association, Chicago, Ill., Sept., 1951.

The desire to experiment with many predictors balanced by the time limit imposed led to a consideration in designing the experiment that there would be a constant battery for trial at both bases, plus certain unique items for each. The unique tests were to be those for which correct answers were available, in general, the aptitude tests. The tests common to both fields were generally those for which keys could not be predetermined, the attitude, interest and personality tests. Table VII below describes the tests used in terms of their time limits, numbers of items and installations at which used.

TABLE VII

TESTS USED IN THE EXPERIMENTAL TEST BATTERY

Number Assigned	Test Name	Time Limit	Number Items	Base at Which Used
I	Supervisory Judgment	30 min.	35 at each Base	Hill and Tinker
II	Personal Preference Inventory	None	80	Hill and Tinker
III	Rosenzweig Picture Problems	None	24	Tinker
IV	Supervisory Problems	None	60	Hill and Tinker
V	Pattern Matching		25	Hill
VI	Personal Preference Inventory	None	98	Hill and Tinker
VII	Number Facility	15 min.	44	Hill
VIII	Personal Preference Inventory	None	37	Hill and Tinker
IX	Biographical Information Blank	30 min.	50	Hill and Tinker
XI	Street-Gestalt			Tinker
XIII	Interest Inventory	15 min.	58	Hill and Tinker

I. DESCRIPTION OF TESTS

Supervisory Judgment. The judgmental aspect of supervisory jobs and the success reported in Chapter II with tests to measure it led to a decision to include a supervisory judgment test in the experimental battery. While an early reference to such a test implied that the manner of responding to the problems reflected a personality rather than a judgmental factor,³ Mandell's subsequent work with tests of this type suggested that the term "judgment" in the test's title might be an accurate one.⁴ In essence, the respondent is presented with a number of facts surrounding situations commonly faced in supervising workers. He is then asked to choose one of five choices which he thinks is the best solution to the problem or reason for it. Difficulty is increased by inclusion of more than one reasonable choice, but only one choice is the best and accorded credit.

Example: In general, the most important advantage of good employee morale is that it results in

- A. high production
- B. decreased work for the supervisor
- C. increased ease in rating workers's efficiency
- D. higher standing for supervisor with management
- E. less desire for wage increases by employees

³ Nancy Timpany, "Assessment for Foremanship," British Journal of Psychology, 38:23-28, September, 1947.

⁴ Milton M. Mandell, "The Administrative Judgment Test," Journal of Applied Psychology, 34:145-147, June, 1950.

Personal Preference Inventory (Tests II, VI, VIII).

To measure varying aspects of personality which were seen in Chapter II to be related to success in supervisory positions, several different types of personality measures were included in the experimental battery. The Personal Preference Inventory comprises parts II, VI and VIII of the battery. These parts are partially patterned after items in the Jurgensen's Classification Inventory which Mandell claimed in oral discussion has shown considerable promise as a base for developmental keying. Test II asked the respondent to select from a triad of personal characteristics which is most irritating and which is least irritating. The triads were developed in the main to ascertain differences in response based on the premise that better supervisors would choose "outgoing" rather than "introverted" responses or "physical appearance" responses, as suggested by Sanford in his studies on authoritarianism and leadership.⁵

Examples: A) always wears bow ties
B) always criticizes people
C) likes to "show off"

1. Which characteristic is most irritating to you?
2. Which characteristic is least irritating to you?

⁵ Fillmore H. Sanford, Authoritarianism and Leadership, (Philadelphia, Pennsylvania:Stephenson Brothers, 1950.

Test VI was included because it was felt that a personality measure with face validity might be a better predictor than one more obscure. Efforts were made again to avoid making one choice more attractive than the others. Triad format was used in asking the respondent to choose the quality of supervisors he thinks most and least important. The theory behind this test was that a projective element could be built into the test which, unaware to the respondent, would reflect his own personality in the responses he chose. The triads included a choice involving a relationship with others, one involving the work itself and a third involving a characteristic of the supervisor as an individual. Based on most studies dealing with basic research in leadership as reported in Chapter II, it was postulated that choices involving relationships with others would be related to success as a supervisor.

Example: A supervisor who
A) plans his work well
B) is personally neat
C) knows his workers

1. Which characteristic would you consider most important in a supervisor?
2. Which characteristic would you consider least important in a supervisor?

Test VIII requires the respondent to choose between two statements about the way people behave. Neither statement of the pair is intended to be more attractive than the other. Respondents are "forced" to choose. This test likewise was felt to be a "projective" test in the sense that an individual's preference might more nearly describe himself and that a pattern related to success as a supervisor might therefrom emerge.

Example: If you had to choose between people who tended to act in the ways described, which person would you choose to include among your friends?

1. A person who
 - A) always "shows off"
 - B) is shy and retiring
2. A Person who
 - A) talks very slowly
 - B) talks very fast

Rosenzweig Picture Problems. Though included in the battery, it was known beforehand that whatever results were disclosed, the Rosenzweig Picture Problems could not immediately be used as a predictor. This is a picture projective test. The respondent is presented with a picture representing some annoying situation. He is then asked to indicate in his own words what response the person in the picture would make to the situation presented. Foreknowledge that this test would not be used was predi-

cated on the fact that a test which could not be scored objectively would not be practical in civilian personnel operations of the Air Force where clinically trained psychologists were not available to interpret results. The test was included in the experimental battery, however, in the hope that from this administration it might be possible to develop multiple choice categories for later research purposes, with the premise that tolerance to petty annoyance is characteristic of the better supervisor.

Supervisory Problems. A further attempt to assess personality, particularly the important qualities of sensitivity to people and open-mindedness was made by inclusion of the Supervisory Problems Test. Some success in using items of this kind by the Civil Service Commission and in preliminary Air Force studies suggested its use in this study. In this test, the respondent is given a list of statements about workers or about people in general and is asked to indicate how he feels about the statement in terms of five possible responses: strongly agree, agree, undecided or uncertain, disagree and strongly disagree. The statements are primarily based on commonsense aspects of everyday psychology, though others are chosen as not having necessarily precise answers. The respondent's choice is believed to reflect his attitudes or beliefs about human relations.

Examples:

1. Most workers today are less efficient than workers 10 years ago.

- (A) strongly agree
- (B) agree
- (C) undecided or uncertain
- (D) disagree
- (E) strongly disagree

2. Poor work habits are probably inherited.

- (A) strongly agree
- (B) agree
- (C) undecided or uncertain
- (D) disagree
- (E) strongly disagree

Aptitude tests. Abstract intelligence was shown in Chapter II to be related to success as a leader. Since abstract intelligence comprises verbal, numerical and spatial sub-factors, and since verbal ability is largely measured by the Supervisory Judgment Test, it was felt that measures of spatial and number ability might add validity to the experimental battery. A Pattern Matching Test (spatial) was included in the experimental battery as a partial measure of intelligence and also for its possible relationship to success as a supervisor in the blue-collar group. In this test the respondent is asked to select from a number of choices showing certain geometric figures, the one set which if properly assembled would result in a given stimulus layout. In addition, a number facility test was included as another factor in intelli-

gence as well as on the theory that a supervisor, in much of his planning, budgeting and evaluating must have a numerical facility - the ability to work rapidly with numerical symbols. The test included merely presents a number of rather simple arithmetic problems and asks the respondent to choose the correct answer from five choices.

Biographical Information blank. While conflicting evidence of the value of biographical data for predicting occupational success exists, the idea that certain background patterns are associated with success is still a tenable one. Biographical data may be merely an extension of application blanks from which predictions of success are frequently made by persons using them as an employment screening tool. The biographical information blanks used in this study include questions of fact and also questions asking the respondent for his own evaluation or description of himself.

Examples:

1. How many living brothers and sisters do you have?
 - A. none
 - B. 1
 - C. 2 - 4
 - D. more than 4

2. Which of the following would you like most about being a supervisor?

- A. the pay
- B. the prestige and authority
- C. the variety of work
- D. the opportunity for advancement

3. How qualified in your line of work do you consider yourself?

- A. exceptional
- B. above average
- C. average
- D. poor

Street-Gestalt. The theory that one requirement of administrators or supervisors is an ability to draw conclusions on the basis of incomplete data prompted the inclusion in the experimental battery of the Street-Gestalt test, a series of incomplete pictures which the respondent is to identify as a whole object. A factor analysis of 29 tests including the Street-Gestalt was reported by Thurstone.⁶ The Street-Gestalt had a high loading in a factor which Thurstone defined as "an ability to form and hold perceptual closure against distraction and ability to shake off one set and take on a new one." In this test, the respondent is given

⁶ Louis L. Thurstone, A Factorial Study of Perception (Chicago, Illinois: University of Chicago Press, 1944) p. 101.

freedom of choice in responding. The test has some projective elements in it (the incomplete pictures reminding one of the Rorschach inkblots), there is one correct answer to each picture, which answer was to be used by the scorer as the key answer. The lack of need for interpretation in the presence of a correct answer made it a practical test to use operationally, in contrast with the Rosenzweig, should its validity be established.

Interest Inventory. Previous research reported in Chapter II indicated that interest items were of value in selecting supervisors and were consequently included in this study, with revised and expanded numbers of questions. The respondent is asked to indicate which of three activities listed in each question he would like most to do and which he would like least.

Example:

- A) develop more efficient office methods
- B) be a personnel technician
- C) write up results of research

1. Which would you like most?
2. Which would you like least?

II. THE TEST SESSION

Installation newspaper publicity brought the attention of those finally selected for the study, its

purpose and the nature of cooperation that might be requested. Supervisors were prepared, therefore, for the testing session. Supervisors were called in groups of twenty to forty to take the experimental battery of tests. Cooperation was excellent. Only one man objected to the test session and he was permitted to leave. The tests were taken in one sitting except for a brief "smoke" break. An insignificant number of literacy problems was encountered and since these tests require reading, those persons were either dismissed or their papers identified and later discarded.

It might be added that criterion data collection was initiated before samples were tested so that as many cases on whom criterion data could not be collected completely could be eliminated from the testing session. Criterion data collection was so time-consuming, however, that in many cases it was done simultaneously with the test administration or after the tests had been administered.

Sessions for administering the experimental battery were in each case conducted by professional research psychologists who were familiar with all aspects of the study. The psychologists represented Headquarters, U. S.

Air Force, Headquarters, Air Materiel Command and either Hill or Tinker Air Force Base. Standardized instructions for administering each test and carefully prescribed time limits developed in advance assured that each test administration session would be conducted under standard, carefully controlled conditions. Testing rooms were designed to provide ample work space, good lighting and ventilation, and freedom from noise and distraction. Each test was preceded by oral and written instructions plus sample exercises so that examinees could ask questions and be checked to assure that they knew what was expected of them. Except for the Street-Gestalt and Rosenzweig tests, questions were answered on IBM answer sheets. Sessions were of four hour duration, with a ten minute break after the second hour.

CHAPTER VI

ANALYSIS OF DATA

I. DESIGN OF ANALYSIS

Procedures for analysis of the data collected were developed around the following requirements:

- (1) Tests included in the battery for which no a priori key was available would be keyed on a portion of the sample and the key thus developed checked on an independent sample. Wherry describes this process.¹ "In certain cases, particularly in those dealing with interests, personality traits, or rating phrases the correct answers are frequently unknown, in which case the technician must first resort to alternative analysis (italics in the original)." And later, ". . . multiple cross-validation is extremely desirable for alternative analysis."
- (2) The combination of tests and their weighting developed on a portion of the sample would be checked on still another independent sample. Thorndike avers that the only test for shrinkage ". . . is to determine intercorrelations and regression weights on one experimental pop-

¹ Robert J. Wherry, Chapter IV, Section 28, "Item Analysis," in Douglas H. Fryer and Edwin R. Henry, Editors, Handbook of Applied Psychology (New York: Rinehart and Company, 1950) pp. 181-182.

ulation, and then apply these to an independent new sample as a test of their effectiveness."² And later, he deems this necessary because ". . . the test that is most valid in a particular sample is so in part because of its genuine validity, but in part because in that sample (italics in the original) that test happened to have a large positive deviation from its true population value."

- (3) Computation of inter-base differences would be disregarded in favor of retaining subsamples of sufficient size for doing the analysis required in (1) and (2) above. With this in mind, both fields would be equally represented in each step of the analysis.
- (4) The Wherry-Gaylord approach would be used in the selection of the battery and designation of weights for sub-tests.³ This method is superior to the Wherry-Dolittle method because it selects tests most predictive automatically and provides weights in integer form which tend to be more stable than those carried out to several decimals.

² Robert L. Thorndike, Personnel Selection (New York: John Wiley and Sons, 1949) p. 204.

³ Robert J. Wherry and Richard H. Gaylord, "Test Selection with Integral Gross Score Weights," Psychometrika, 11:173-183, September, 1946.

- (5) Occupational differences would be investigated, if feasible.
- (6) Several methods would be used in developing keys for personality, interest and attitude tests and each checked as described above.
- (7) The hypothesis would be rejected that the obtained multiple correlation could have resulted from chance sampling in which the true multiple correlation is zero. The test of validity of any proposed battery should consist of a demonstration that the obtained r of battery scores against the criterion of success in an independent sample is statistically significant irrespective of its absolute magnitude; that is, that the obtained r exceeds the .61 level of confidence in terms of the deviation from zero.
- (8) The hypothesis that the true multiple correlation would be equal to or greater than some value high enough for practical predictive efficiency would be accepted. The success of the efforts would be measured in fiducial terms rather than in terms of individual prediction.

Grouping of data. In terms of the above requirements, the subsamples for keying, battery selection and final checking would be developed as follows:

- (1) Subsample I - One-third cases from Hill and Tinker Air Force Bases combined, selected in random fashion. Used for developing tentative keys on personality tests.
- (2) Subsample II - One-third cases from Hill and Tinker Air Force Bases combined, selected in random fashion, to be used for checking the keys derived from subsample I and developing a tentative battery of final tests.
- (3) Subsample III - One-third cases remaining to be used for checking the battery and weights.

This design provides for the very necessary feature of replication or cross-validation. Johnson states the condition that replication assures precision and ". . . is the sole source of the estimate of error."⁴ The existence of error is considered by Gullicksen to be one of the most basic assumptions of all measurement, of which test theory is a special case.⁵ Cureton describes the procedure adopted in this study in detail.⁶ We need one sample for item se-

⁴ Palmer O. Johnson, Statistical Methods in Research, (New York: Prentice-Hall, Inc., 1949) pp. 280-281.

⁵ Harold Gullicksen, Theory of Mental Tests, (New York: John Wiley and Sons, Inc., 1950) p. 4.

⁶ Edward E. Cureton in E. F. Lindquist, Editor, Educational Measurement, (Washington, D. C.: Amer. Council on Education, 1951) p. 693.

lection and item weighting. We need another for test selection and test weighting. . . . We need a third sample to determine the predictive power or validity of the battery. Every time we violate one of these rules, we increase spuriously the apparent validity or predictive power of our test battery." This procedure increases our confidence that the results obtained may be generalized from this sample to the population of supervisors as a whole.

Use of stanines. To simplify data handling, all raw scores and criterion scores were converted to stanines which are standard scores with a mean of 5 and a standard deviation of 1.9. By converting all two and three digit scores to stanines rounded off as integers, reduction in calculating machine time and errors was effected. This step is warranted since no test or rating is so precise that fine gradations in scores represent real differences in ability.

II. KEYING PERSONALITY TESTS

The three sections of the Personal Preference Inventory, the Biographical Information Blank, and the Interest Inventory required special procedures for development of unique keys which would best differentiate the better from the poorer supervisor in terms of the criterion. Two approaches to keying these tests were applied

and the second was in fact based upon hypotheses suggested while engaged in the first.

Empirical keying. The first approach taken with the test material not having a priori "correct" answers was to separate the sample into two groups on the basis of the criterion: a "high" group consisting of subjects in stanines 6 to 9, and a "low" group made up of those in stanines 1 to 4. Percentages of persons in each group choosing each response were calculated and compared. Lawshe describes this procedure clearly.⁷ "All of the items that logically appear to measure a given area . . . are identified, and the test papers of a trial group are scored on these items only. The proportion of each of two criterion groups is checked and if they are the same, the item is dropped. This process is said to employ the criterion of internal consistency." In general, items were selected for keying if

- (1) one particular choice was selected by a more or less significantly greater proportion of the more successful supervisors than the less successful.
- (2) the choice was selected by a reasonable number of persons so that occurrence of that event was not as likely to be a sampling artifact.

⁷ Charles H. Lawshe, Jr., Principles of Personnel Testing, (New York: McGraw Hill Book Co., 1948) p. 89.

- (3) the item met a standard of rationality if thus keyed - i.e., the item could be explained as having some rational explanation or logical relationship to success as a supervisor. In a sense this might be called a "job analysis" test.

As the analysis progressed, two different standards of empirical keying were employed: first, where an alternative was keyed if the difference was significant at the 50% level of confidence; later where a standard of discrimination between the two criterion groups significant at the 20% level of confidence was applied.

"Unconventional" keying. As the empirical keying process proceeded, it became evident that choices simply selected by fewer people tended to differentiate between the better and poorer criterion groups. The influence of this hypothesis worked in a negative, or opposite direction from the empirical key. It identified the less successful. Inspection of the data suggested the desirability of applying this scoring approach as a second key to the test score data. Because it was based on selection of a choice by few respondents, it was termed an "unconventional" key. The term has no value judgment; it merely refers to the rarity with which an alternative was chosen. A possible theoretical explanation for this approach lies in the fact that a certain stereotype of behavior and thinking has been ac-

cepted as being associated with good supervision and that deviations from this stereotype are thought to be unsuccessful. If this hypothesis is correct, it implies the perpetuation of whatever we have now in future selections for supervisory positions. It is not for this report to judge the relative social or economic advantages of this result. To some, the spark of genius which may exist in one individual, however curious he might otherwise appear or behave, is worth the possible difficulties which may have to be endured. To others, the feeling of security of a relatively homogeneous work-force at the supervisory level is comforting or even demanded. Whatever the feeling, the methodology of psychological test development is geared toward perpetuating the qualities deemed and tested as successful. The improvements wrought are more in matter of degree than of kind. Though design of studies has not advanced to where tests can be developed in other ways, it does ignore the obvious fallacy that because all A is B, all B is necessarily A.¹ We say that all supervisors with a given set of qualities are successful; we may not logically conclude that as a consequence all successful supervisors have this set of qualities. It may merely be that we have not sampled that variety. Or if we have, the lack of complete agreement has been partially responsible for something less than perfect validity in our selection

¹ Morris R. Cohen and E. Nagel, An Introduction to Logic and Scientific Method. (New York: Harcourt, Brace and Co., Inc., 1934) pp. 60-62.

instruments. The truest test into ascertaining success of persons with wide variance in qualities would be a purposeful employment of all persons with foreknowledge of their differences, placing them in comparable positions with carefully controlled and similar circumstances and then observing them in their performance. This is something industry is not always willing to do, and one can appreciate their reasons therefor.

It is the method of psychological research, at any rate, to identify those qualities common to the sample which are associated with successful job performance. The "unconventional" keying approach is based on the hypothesis that there is a source of valid variance which might be ignored by the large-response approach of pure empirical keying. This hypothesis was verified as the "unconventional" keying process advanced, but the ability to differentiate naturally occurred less when the direction of difference was ignored. Two approaches were therefore used: one in which the "unconventional" key was applied, regardless of whether or not it was in error in direction of discrimination; the other in which only those items were keyed "unconventional" if (a) they were selected by fewer than 5, 8 or 10 percent of the population, depending upon the number of alternatives in the item, and (b) they discriminated success-

fully in the correct direction. In either case, the standard of rationality applied to the empirical key was likewise applied to the "unconventional" key. We were fortunate in having available for this rather laborious process of item analysis, the facilities of an IBM machine room. It was possible to have answers punched on cards for all individuals in the sample, and counts made of all choices taken by persons in various criterion groups.

III. STEPS LEADING TO MOST PREDICTIVE PERSONALITY BATTERY

Testing the keys. The first step in the analysis was devoted to testing both the empirical and the "unconventional" keys. For this run, fields were combined and occupations were combined. The empirical key was based on no specific probability standard, nor was a discrimination standard applied with the "unconventional" key. An arbitrary yield of ten percent of the items, however, was specified. The Pearson product-moment correlations were computed for these keys singly and also combined in terms of the total test scores. The results are shown in Table VIII below. It should be noted that these keys were derived on subsample I and that the reported results are based on results obtaining in subsample II, a completely independent sample. It is interesting further

to note that the "unconventional" key behaves as expected, by producing correlations with negative signs. By subtracting the "unconventional" scores from the empirical scores, a third score for each individual was computed which, it may be observed, does as well as or better than either one alone in four of the six tests as well as the total.

TABLE VIII
CORRELATIONS OF THE PERSONALITY SUB-TESTS
WITH THE CRITERION

	Tests						
	II	IV	VI	VIII	IX	XIII	Total
Empirical key- 50% standard	.11	.11	.15	.05	.12	.13	.21
Unconventional key. Random standard	-.04	-.09	-.10	-.14	-.11	-.17	-.20
Empirical minus Unconventional	.04	.12	.16	.10	.15	.17	.23
Note: N = 153. Fields and occupations combined.							

By combining only those tests which in any row provided a correlation of .14 or better, namely Tests VI, VIII, IX and XIII, the correlation for the total of the four tests when scored "empirical minus 'unconventional'" was .25, somewhat better than the total of all six tests when treated in the same manner. As a further replicative check, these same four tests, VI, VIII, IX and XIII, were applied with the "empirical minus 'unconventional'" scoring procedure to subsample III to yield a correlation of .20, showing remarkably little shrinkage.

When this same key for the same four tests was applied to specific occupational groups in subsample III, the results were for

Professional supervisors	.45; N = 17
Clerical supervisors	.21; N = 30
Blue-collar supervisors	.17; N = 108

Further inspection of Table VIII suggested selecting and combining tests with the highest correlations with the criterion, disregarding the factor of consistency in manner of keying. This led to selection of Test VI keyed empirically, test VIII keyed "unconventionally," and Tests IX and XIII keyed "empirical minus 'unconventional.'" If E were to represent empirical and U unconventional, the score based on the most predictive tests would be as follows:

$$VI E - VIII U + (IX E - IX U) + (XIII E - XIII U)$$

This key when applied to subsample III improved the correlation, yielding a coefficient of .24, significant at the 1 % level of confidence.

Further considerations in keying. One the basis of the foregoing preliminary results, consideration was given to:

- (1) Occupational keying. Since the obtained relationships of the personality tests revealed occupational differences, the sample combined as to occupation may have been masking

this fact and producing lower correlations. Because of the satisfactory size of the blue-collar group, it was decided that that occupational sample would be separated from the other groups and keys derived specifically for it.

- (2) A more stringent requirement for empirical keying to see what effect this might have in sharpening the relationship between tests and criterion. It was decided that a standard would be placed on the empirical key requiring items to discriminate minimally at the 20 % level of confidence before being included for scoring. This key will be designated as " $p \geq .20$ " probability equal to or less than 20% that the occurrence is due to factors of chance. The empirical key used in earlier steps will be referred to as "50% standard" meaning that items were included if the direction of proportions of the respondent population choosing the alternative indicated a relationship to the criterion.
- (3) A more stringent requirement for the "unconventional" keying to see what effect this might have in sharpening relationships between tests and criterion. Whereas earlier all choices answered by a small number of respondents were

keyed, in many cases the key erred in direction of favoring the poorer people. In the next attempt, it was decided that only those choices would receive an "unconventional" label which, in addition to being answered by a small proportion of the respondents, also favored validity by discriminating in the direction of identifying the less successful supervisor in terms of the criterion. This "unconventional" key would be identified as "hits" while the former "unconventional" key would be identified as "random standard."

- (4) Giving a last minute review to all items to be keyed in any of the four ways to determine whether or not all keyed items had a rational explanation, "face" validity as it is sometimes called, or possible unsatisfactory public relations effects. It was decided that such items would be eliminated before final analysis.
- (5) Discarding negative keying of empirically keyed items which would warrant a negative score. Scanning revealed few such items, and it was felt that the poor public relations effects of screening out persons

on this basis together with the added inconvenience of plus and minus keys for empirical scoring would more than balance what small contribution to validity the additional items might make.

Four keys applied to blue-collar group only.

Breaking out the skilled group only and applying now the four keys to the six personality tests: empirical, $p \geq .20$; empirical, 50% standard; "unconventional" hits; and "unconventional" random standard, the following results emerged:

TABLE IX

CORRELATIONS OF PERSONALITY TESTS WITH THE CRITERION

	Tests						Total
	II	IV	VI	VIII	IX	XIII	
Empirical, 50% standard	.21	.00	.13	.01	.25	.15	.17
Empirical, $p \geq .20$	.22	.03	.01	-.02	.07	.17	.15
"Unconventional" random standard	-.09	-.03	-.13	-.09	-.29	-.15	-.26
"Unconventional" hits	-.12	-.07	-.19	-.07	-.22	-.15	-.23
Note: N = 108. Blue-collar group only. Table based on key developed from subsample I and applied to subsample II.							

Multiple correlational analysis of personality tests. Up to this point, test selection has been based on inspection, identifying those tests with higher correlations with the criterion and eliminating those with lower correlations. The total test score has represented merely the summation of raw scores of selected subtests. It was felt at this point that by application of

multiple correlational techniques to a battery of judiciously selected tests, increased predictive power would be built into the "personality" test battery. In Thorndike's words, "the multiple correlation serves as an index of the degree to which a test battery is successful in predicting a criterion."⁹ Based on the coefficients in Table IX, selecting arbitrarily those tests with coefficients of .12 or higher, the following battery was chosen:

	<u>Tests</u>
Empirical, 50% standard	II, VI, IX, XIII
Empirical, $p \geq .20$	II, XIII
"Unconventional" random	IX
"Unconventional" hits	II, VI, IX, XIII

Applying the Wherry-Gaylord test selection method to this battery of eleven tests with the sample restricted to the blue-collar group of subsample II, representing both fields, the multiple correlation coefficient was found to be .43 with five tests weighted as follows:¹⁰

<u>Weight</u>	<u>Test</u>	<u>Keying Method</u>
1	IX	Empirical, 50% standard
1	II	Empirical, $p \geq .20$
3	IX	"Unconventional" random
1	VI	"Unconventional" hits
2	II	"Unconventional" hits

A multiple correlation of .43 is a reasonably good relationship, particularly in connection with "personal-

⁹ Robert L. Thorndike, Personnel Selection, (New York: John Wiley and Sons, Inc., 1949) p. 189.

¹⁰ Robert J. Wherry and Richard H. Gaylord, "Test Selection in Integral Cross Score Weights," Psychometrika, 11:173-183, September, 1946.

ity" tests. The importance of retaining a sample for cross-validation purposes, however, is clearly seen from the following. The aforementioned battery of five tests derived from the Wherry-Gaylord integral test selection process, was applied to subsample III, composed as subsample II was composed, of only blue-collar supervisors representing both bases. A disappointing correlation, $R = .15$, was obtained. When the derived weights were disregarded, and tests permitted to weight themselves on the basis of number of items correctly answered, that is, the total raw score would represent the score unweighted, the correlation was boosted to $.22$.

Cause of shrinkage of correlation. Experience in aptitude test work reflects the lack of stability and consequently confidence that may be placed either in the coefficient of correlation or in the weights for the sub-tests in a battery. In aptitude test work it has been found rather characteristically that tests of a certain type will predict success in a given field of work, but that the correlation coefficient may vary within extremely wide ranges. Guilford points out that " . . . the validity as indicated by a multiple R applies strictly to the group of subjects from which the regression equation was obtained. When the test battery is applied to a new group, there is typically a

shrinkage (italics in the original) in the size of R and the regression weights may vary from those obtained earlier."¹¹ Cureton has most vocally flailed the use of multiple regression coefficients in psychological research.¹² "(a) Only in exceptional cases are the multiple regression coefficients of a criterion score upon a set of test scores the proper weights to give the test scores in order to predict or estimate the criterion scores. (b) When the statistics from a given sample have been used to determine the test score weights, the estimate of the aggregate or multiple correlation of the tests with the criterion, as computed from the data of that sample, is not an estimate (italics in the original) of the predictive power or validity of the battery." Wherry, in oral discussions, has referenced a study completed by one of his graduate students which concluded that if test intercorrelations are reasonably high and sample sizes are under 200, unit weighting serves to yield less shrinkage on independent samples than either beta weights derived from Wherry-Dolittle multiple correlation methods or integral weights from Wherry-Gaylord test selection methods. Where test

¹¹ Joy P. Guilford, Psychometric Methods, (New York: McGraw Hill Book Company, Inc., 1936) p. 426.

¹² Edward E. Cureton in E. F. Lindquist, Editor, Educational Measurement, (Washington, D. C.: American Council on Education, 1951) p. 690.

intercorrelations are near zero, however, and sample sizes are over 200, beta weights or Wherry-Gaylord weights give better predictions on independent samples.¹³ Additional tests of this phenomenon accomplished with other data available to the author, confirm this statement. Whether or not the weights remain stable by either previous method, the multiple correlation technique is still expected to provide the tests which in combination best predict the criterion.

Rational Test Selection and Unit Weighting. Recalling that unit weighting of tests selected by the Wherry-Gaylord method resulted in correlation with the criterion of only .22, it was decided to combine the tests into batteries on a rational basis, with no attention to possible differential weight assignment to the sub-tests. Instead, tests were selected for inclusion in the battery on the basis of adjudged homogeneity within the tests and adjudged heterogeneity of tests within the battery. Tests IV and VIII would be excluded because of different format used in those and because statistical results thus far did not confirm their predictive value with the criterion. The batteries thus developed were

¹³ Discussions held 29 and 30 January, 1953 with Dr. Robert J. Wherry, Professor of Psychology, Ohio State University while serving as consultant to Air Materiel Command in Washington, D. C.

$$\begin{aligned} E_1 &= U_1 \\ E_2 &= U_2 \\ E_1 &= U_2 \end{aligned}$$

where

$$\begin{aligned} E_1 &= \text{empirical, 50\% standard, tests II, VI, IX and XIII} \\ E_2 &= \text{empirical, } p \geq .20, \text{ Tests II and XIII plus empirical, 50\% standard, Tests VI and IX} \\ U_1 &= \text{"unconventional" hits, Tests II, VI, IX and XIII} \\ U_2 &= \text{"unconventional" hits, Tests II, VI, XIII plus "unconventional" random, Test IX} \end{aligned}$$

Extracting the blue-collar group for subsample III, an independent sample representing both fields, the results were as follows:

$$\begin{aligned} E_1 - U_1; \quad r &= .33 \\ E_2 - U_2; \quad r &= .27 \\ E_1 - U_2; \quad r &= .32 \end{aligned}$$

While the difference between the $E_1 - U_1$ correlation and the $E_1 - U_2$ correlation is quite small, the $E_1 - U_2$ battery was chosen as the final battery because the single sub-test that differed in the two formulae had more keyed items when no standard was applied and consequently had a wider range of scores. Applying the keys derived on the blue-collar supervisors to clerical and professional groups, the results were

$$\begin{aligned} \text{Clerical supervisors,} \quad r &= .36 \\ \text{Professional supervisors, } r &= .15 \end{aligned}$$

Face validity of questions which were satisfactory with blue-collar employees was not always satisfactory for clerical and professional supervisors so that item changes would be required if the test was to be used with those groups operationally. Such changes would of course require additional validation to test their effectiveness. Furthermore, sample sizes of clerical and professional supervisors were somewhat small to permit generalization with confidence to all supervisors in those categories. The resultant battery is to be recommended, therefore, only for blue-collar employees.

Data for specific tests in the final personality battery. The intercorrelations of the tests in the personality batteries ranged generally from .2 to .4 with Test IX showing itself to be the best test in terms of highest correlations with the criterion and lowest intercorrelations with other tests in the battery. Table X below depicts the frequency of intercorrelations achieved for the various personality tests in the final battery.

TABLE X

INTERCORRELATIONS OF TESTS IN FINAL PERSONALITY BATTERY
IN TERMS OF FREQUENCY

Intercorrelations	All tests	Test IX only	All minus Test IX
.71 - .60	1	1	1
.61 - .70	1	-	1
.51 - .60	-	-	-
.41 - .50	3	-	3
.31 - .40	3	-	3
.21 - .30	3	-	3
.11 - .20	0	4	0
.00 - .10	7	2	5
	<u>20</u>	<u>13</u>	<u>15</u>

Except for Test IX, the intercorrelations suggest that entirely discrete characteristics are not being measured by the sub-tests. In Table XI may be found the correlations of the sub-tests in the final personality battery for 108 blue-collar employees in subsample III, representing both fields. The means and standard deviations of the blue-collar supervisors of subsample III on the aforelisted personality tests at each field will be of interest in terms of the expected consistency in score ranges which may be expected by the Air Force at diverse Air Force Bases.

TABLE XI

CORRELATIONS OF TESTS IN FINAL PERSONALITY BATTERY
WITH THE CRITERION

	Test	Correlation
Empirical, 50% standard	Test II	.21
	VI	.13
	IX	.25
	XIII	.25
"Unconventional," random standard	IX	-.29
"Unconventional," hits	II	-.12
	VI	-.19
	XIII	-.15
Note: N = 108. Blue-collar employees from two Air Force Bases.		

TABLE XII

COMPARISONS BETWEEN AIR FORCE BASES IN SCORES ON THE PERSONALITY TESTS

	Hill AFB	Tinker AFB	Total
Mean	55.2	54.4	54.8
Standard Deviation	10.5	13.0	11.8
Number	49	59	108.

Number of personality items keyed. Table XIII below indicates the number of choices actually keyed on the

final "personality" battery compared with the total number of choices in the original experimental battery. It is gratifying to observe that 25% of the possible responses to have been keyed have actually been discriminating enough to warrant keying. Forty-six percent of the items in the original test have been retained because of their discriminability. This is a phenomenal retention figure since it is based on the ability of the item to discriminate on an entirely independent sample with no element of "boot-strapping" involved.

NUMBER OF RESPONSES KEYED IN FINAL PERSONALITY BATTERY

Test	II	VI	IX	XIII	Total
Number of Responses in Original	240	294	200	174	908
Number of Items in Original	80	98	50	58	286
Responses Empirical, 50% standard	36	36	24	32	128
Responses "Unconventional," random standard	-	-	38	-	38
Responses "Unconventional," Hits	12	18	-	17	47
Total Response Keyed	48	54	62	49	213
Total Items Keyed	32	32	42	26	132

IV. THE APTITUDE TESTS

The aptitude tests in the experimental battery were given at only one or another of the two bases because of practical considerations described earlier. In addition, it was felt that experience would lend confidence to the stability of the aptitude test results in terms of the relationships they would maintain. Except for the Supervisory Judgment Test and the Street-Gestalt

Test, aptitude tests showed little relationship with the criterion. This was not as disappointing as it might seem. For one thing, in obtaining ratings, emphasis was placed on disregarding technical competence and attending only to supervisory competence, which may be considered largely a personality and intellectual configuration. Furthermore, application of aptitude tests to specific populations of supervisors might not be expected to yield much of a relationship because of their expected homogeneity with reference to that aptitude.

Number Facility. Number Facility used at Hill Air Force Base correlated .13 with the criterion for the total blue-collar group of 150 cases; .34 with the clerical group of 30 cases; .05 with 37 professional supervisors and .09 with the total group. Though the numerical factors shows some relationship in the clerical fields, with only 30 cases the conclusions must be regarded as tentative.

Pattern Matching. Used only with blue-collar employees at Hill Air Force Base, Pattern-Matching gave a validity coefficient of only .11 and the test was consequently dropped from further consideration.

Street-Gestalt. Used only at Tinker Air Force Base, Street-Gestalt correlated .22 with 178 blue-collar

supervisors, significant at the 5% level of confidence and suggesting its inclusion in the final battery.

Supervisory Judgment Test. At Hill AFB, the 35 supervisory judgment items used correlated .30 with the criterion for blue-collar supervisors in subsample III, while at Tinker, another 35 items correlated .22, suggesting value to be derived by using these tests in the final battery.

Combined prediction. To test the effect of Supervisory Judgment (Tinker) plus Street-Gestalt as combined predictors, a Cherry Gaylord test selection procedure was accomplished on subsample I at Tinker AFB:

r Supervisory Judgment vs criterion	.30
r Street-Gestalt vs criterion	.22
R Supervisory Judgment plus Street-Gestalt vs criterion	.31

V. THE FINAL OVERALL BATTERY

Addition of Supervisory Judgment to the final "personality" battery raised prediction of the criterion from .32 ("personality" tests alone) to .39. Addition of Street-Gestalt at Tinker raised the correlation to .37. Addition of Street-Gestalt and Supervisory Judgment raised the correlation to .38. In spite of the fact that the two additional tests did

not improve the correlation over addition of either one alone, recommendation is made to use both tests. In the first place, the smaller number of items in the aptitude as compared with the "personality" tests means that they will have less weight in the final battery. Secondly, the non-verbal features of the Street-Bestalt test might make it more acceptable to blue-collar workers. The final battery, therefore, would be

$$Z_1 = Z_2 + 30 + 35$$

Accepting the correlation of .36 based on this combination as an uncontaminated correlation from an independent sample, a correction for attenuation was applied. This correction is justified because of the restricted range of ability one expects to find in using an employed sample for research purposes. It is computed from the formula

$$r_{\text{cor}} = \frac{r_{xy}}{\sqrt{r_{yy}}}, \text{ where}$$

r_{xy} is the validity coefficient

r_{yy} is the reliability of the criterion

In this case, the correlation corrected for attenuation becomes .49, the maximum validity which could be expected from use of this battery. Final statistical results are summarized in Table XIV below.

TABLE XIV

MATRIX OF INTERCORRELATIONS OF FINAL BATTERY TESTS
AND COMBINED CRITERION

	Person- ality	Supervisory Judgment	Street- Gestalt	Cri- ter- ion
Personality		.54	.57	.32
Supervisory Judgment			.46	.30
Street-Gestalt				.22
R uncorrected, .38 R corrected .49 Note: Based on subsample III, blue-collar em- ployees. N = 59.				

VI. CLERICAL STEPS IN PROCESSING THE DATA

The reader may be interested in the manner in which clerical processing of the analysis described in this chapter took place. While all data were collected in the field, all processing of data was accomplished at a central location in Washington, D.C. Test papers, answer sheets and criterion data, both superior and subordinate, were mailed to that location. The clerical operations were performed by trained statistical clerks who did their calculations on bases established by professional psychologists.

Cards 5 x 8 in size sectioned into cells to simplify clerical operations were prepared for recording data for each individual sample. The numbers 1 to 9 appeared across the card once on each half. Two lettered columns of ten letters each, one column preceding each series of digits, also appeared on the card. Thus, the card provided a simple means of recording data for twenty variables in terms of stanine scores (1 to 9). In addition, the name of the respondent and his location were recorded on the card.

The aptitude tests for which predetermined keys were available, were first scored with stencil keys placed over the IBM answer sheets. Since the correct answers were punched on the key, a respondent's answer, if correct, would be visible as a black mark on his answer sheet. His score then would be simply the number of black marks showing through the key. The scoring was in each case checked by a second clerk. The raw scores were converted into stanine scores according to prearranged methods and the resultant stanine scores were recorded on the 5 x 8 cards. The criterion evaluations for each individual in the sample was then computed as the average of the mean superior plus the mean subordinate rating and recorded on the card.

The total deck of cards representing the sample from both Hill and Tinker Air Force Bases was then shuffled and subdivided into three piled on a random basis, representing the three subsamples described earlier in this chapter. For subsample I, answers for each respondent to the "personality" tests were recorded from the answer sheets onto IBM master sheets. Also coded and recorded were the respondent's age, sex and criterion data. IBM analysis provided runs of all items against the criterion variate. From this, key was prepared by professional psychologists in terms of the empirical and "unconventional" keying methods described earlier.

The clerks then scored subsample II with the keys thus prepared. Correlations were then computed and checked for each test, using each key against the criterion for all occupational groups combined. A table was prepared to record the correlations. The clerks used calculating machines to obtain the correlations in a way using no judgment on their part. After decision to limit the sample to blue-collar employees and to impose differing standards of keying, the processes were repeated.

The battery of tests most predictive of the criterion was developed from subsample II, first using the Wherry-Gaylord approach. Again the clerks followed routine procedures and by means of calculating machines, arrived at the tests and weights most predictive of the criterion. These weights were then applied to subsample III, after the papers had been scored and checked with the keys of selected tests. Correlations were now computed on subsample III papers. Because the results were disappointing, a slightly different combination of tests was selected on a rational basis from data derived on subsample II, and weights were disregarded. Clerks then applied this to subsample III and from correlations with the criterion, obtained what finally was shown to be the recommended test battery.

VII. AVAILABILITY OF TESTS IN THE FINAL BATTERY

Because of the loss of value of test materials arising from their becoming common knowledge and to safeguard the security of the test items and answers, the author regrets that the final tests and keys as published

by the Air Force are not available to organizations outside the Department of the Air Force.

Comparable items to those included in the Supervisory Judgment and Street-Gestalt tests are available to federal agencies from the Test Development Section, Placement and Examining Division, U. S. Civil Service Commission, Washington 25, D. C. The interest items were adapted from the Kuder Preference Record available from Science Research Associates, Chicago, Illinois and then specifically keyed on the Air Force population. Portions of the Personal Preference Inventories were adapted from the Clifford Jurgensen Classification Inventory available from the author at the Minneapolis Gas Light Company, Minneapolis, Minnesota and specifically keyed on Air Force samples. The remaining materials were developed by the Air Force for use in the research reported herein.

CHAPTER VII

PROBLEMS IN USE OF THE BATTERY

In the ordinary course of a research paper, it is perhaps not usual to continue beyond the research conclusions, and to describe the application of the research results, if any is made. The fact that the research reported herein however was accomplished in an administrative setting, was of an applied nature, and was intended from its origin for a specific use within a federal department suggests that the aspects of integration of the research results into the existing structure of personnel administration merit consideration in this dissertation.

If a test is going to be used in a large federal department like the Department of the Air Force, with some 130 field units, some having trained psychologists, others not, representing small and immense organizations, missions of various kinds, and every geographical section of the country with differing educational, economic and religious groups, it is obvious that more planning for its use must be done than the mere printing and distribution of the test.

I. STANDARDIZED DIRECTIONS

So that the test may be administered and scored by persons without formal psychological training, preferably clerical personnel, directions for administration and scoring of the tests in the battery must be explicit and detailed. It is of course of utmost importance that all persons being examined be given identical opportunity to succeed in the competition. Standardized instructions for administration and scoring serve at least in part to insure the realization of that principle. The directions, following those prepared for other civilian Air Force test batteries, will include instructions on what materials to have gathered together in advance of the test session, how to put the group at ease, what materials to distribute, word-for-word instructions to be read aloud, how to handle the practice exercise sessions, and specific timing for each sub-test. This document becomes part of the Personnel Research Test Series of the Air Force and, like the tests, will be stocked and distributed upon requisition from accredited Test Control Officers from two central locations.¹ Scoring instructions are equally detailed and special hand-scoring stencil keys are available which

¹Air Force Regulation 9-3, 30 October, 1952.

enable a test scorer to compute the score for a part of the test merely by counting the marks on IBM answer sheets made by respondents which show through the holes representing correct answers punched on the stencil key.

II. DIRECTIONS FOR INTERPRETATION

Since the interpretation of test results and their application will be accomplished by placement advisers, in many cases without formal psychological training, the directions for interpretation must be as explicit as those for test administration and scoring. Tables of norms will be prepared and included in the Air Force Civilian Personnel Manual.² The norms will describe test scores in terms of the criterion ratings they most generally represent in the experimental population. For example, a final score may be described as "within the range of scores received by those in the experimental population who were generally considered by their supervisors and subordinates to be 'typically effective' supervisors". This information allows the placement adviser to translate a test score into meaningful terms of job performance with which both he and the operating official who ultimately makes the selection are familiar.

²Air Force Civilian Personnel Manual, AFM 40-1, Chapter P.5, Placement, Section 5, Personnel Measurement.

III. USE OF NON-TEST VARIABLES

It is contrary to Air Force policy to utilize test scores as a sole basis for taking an in-service personnel action. All other information, including education, experience, evaluation of previous job performance and the like, is to be considered along with the test score.

IV. LOCAL ADMINISTRATIVE DETERMINATIONS

Because of the varied circumstances surrounding each base in terms of familiarity personnel and operating officials have had previously with use of tests, it was considered undesirable to impose the test as a mandatory requirement at all installations. Command Headquarters were to work with any or all of their installations to determine how the supervisory selection battery could best be used. A pamphlet developed by the Civil Service Commission on selection of supervisors suggested a list of questions to be answered after consideration by any organization establishing a supervisory selection program.³ These questions were made available to all installations for prior consideration before setting up programs of their own. In brief, these and others are:

³U.S. Civil Service Commission, Selecting Supervisors. (Washington, DC: U.S. Government Printing Office, 1951), pp. 23-25.

(1) Will taking the test be a prerequisite to promotion to a supervisory position? If so, must a candidate pass the test? If he fails, is he automatically disqualified or may he be selected in spite of his test score?

(2) Shall a roster of successful candidates be developed and appointments made in rank order or shall selections be made from broad categories, such as A, B, C, Ineligible? Or shall selection be made merely on the basis of having passed, with no preference given for higher test scores?

(3) What ~~qualification~~ requirements will be imposed before an employee is eligible to take the test? What job areas will be included? Will only employees in the level of job just below a supervisory one be permitted to take the test?

(4) Will employees in appropriate job areas be examined only at the time supervisory vacancies occur or will all eligible employees be examined in a concentrated period of time, their scores then to be available in personnel records whenever a vacancy does arise?

(5) How often will the test be administered, balancing security of test materials and possible practice advantages with the need for examining recent appointments and the desirability of giving previous failures a chance at re-testing?

(6) What weight will the test be given in making final selections? Will operating officials be given only a panel of qualified eligibles or specific test score information about all persons thereon?

(7) Will employees be notified of test results? Will appeals on the basis of test scores be permitted?

(8) Will all organizations or occupational groups on the base be covered, or only those with known or anticipated turnover or expansion?

(9) Will local employee groups be contacted to gain their support for this program?

(10) How will the new requirement be "sold" to both management and employees?

Although preliminary discussions were held with interested Command and Headquarters, U.S. Air Force groups to attempt to answer these questions in advance for the purpose of issuing standardized instructions, it soon became apparent that the only practicable solution was to issue the test materials and to impose as few administrative requirements

as possible. Consequently, except for efforts to assure appropriate technical use of the materials and their security, the tests were made available to installations with only a list of considerations to enable them best to plan the institution of the test in the civilian personnel program. After a reasonable lapse of time, it is intended that installations will report on the decisions made with respect to these optional administrative arrangements and the problems and relative success attendant upon use of the test in that manner. On the basis of these responses, it may later be possible to draw up a suitable standardized set of instructions for all installations.

V. PUBLIC RELATIONS CONSIDERATIONS

Management. Management's decision to use tests is rightly one concerned with costs. The benefits to be derived must offset the additional costs of using them. The costs of using the tests may be readily computed in terms of non-productive worker time involved in submitting to the examination, plus the relatively minor amount of personnel office time in scheduling test sessions, and in administering and scoring the tests. The benefits to be accrued in

management's terms are stated in terms of improved supervisory effectiveness which it is anticipated results in increased worker productivity, reduced absenteeism and turnover, and improved employee morale. The validity coefficient expresses the degree to which the test predicts management's own evaluations of their present supervisory workforce. By setting a cut-off score at any point, knowledge of the improvement the test can make over previous selection methods can be computed. Management can better understand a graphic presentation of the validity of the test, however. Most simply, the validity coefficient can be expressed as the ability of the proposed test battery to have screened out, if used, a given number of "duds" now on the payroll.

Taylor-Russell Tables. Another method of expressing the efficiency of a test battery requires decisions or judgments on the part of management using the tests in the selection program.⁴ By bringing management into the program for using the tests as opposed to the psychologist alone making these decisions, not only familiarizes management with personnel research and management but also discloses that suc-

⁴Joseph Tiffin, Industrial Psychology (New York: Prentice-Hall, Inc., 1952), pp. 79-82 and App. B.

cess or failure of the test in part rests with management itself. This method involves use of the Taylor-Russell tables which indicate the proportion who will be satisfactory among those selected for given values of the proportion of present employees considered satisfactory, the selection ratio and the validity of the test to be used. Of the three considerations, the psychologist can provide only the validity coefficient. Management itself must judge the portion of present staff considered satisfactory and decide on selection ratio, taking into account the all-important factor of labor market supply and demand. Management will then be able to decide whether to take only persons with highest test scores or accept also those lower in the test score range. Precise knowledge of the proportion of present workforce considered satisfactory is seldom available, though error in this respect is not at all damaging, merely giving a less accurate account of the efficiency of the test battery in improving over present selection methods. Table XV describes the predictive efficiency of the proposed supervisory selection battery based on the validity coefficient of .49 (as corrected for attenuation and rounded off to .50 for ease in use of the Taylor-Russell tables).

TABLE XV

**INCREASE IN PREDICTIVE EFFICIENCY
BY USE OF PROPOSED TESTING BATTERY**

Proportion of Present Employees Considered Satisfactory	Expected Proportion of Satisfac- tory Employees by Application of Various Cut-off Scores ($r = .50$)				
	39-above (A) 20%	52-above (B) 40%	66-above (C) 70%	47-above (D) 95%	46-below (E) 100%
.05	.15	.16	.07	.05	.05
.10	.25	.19	.13	.11	.10
.20	.44	.35	.26	.21	.20
.30	.58	.48	.38	.31	.30
.40	.69	.60	.49	.42	.40
.50	.78	.70	.60	.52	.50
.60	.86	.79	.70	.62	.60
.70	.91	.87	.80	.72	.70
.80	.96	.93	.88	.82	.80
.90	.99	.97	.95	.92	.90

Table XV clearly demonstrates the value of accepting candidates only in the A group with scores of 89 and over rather than the succeeding groups. If management appraises itself as having 50% of its present supervisors as satisfactory, an increase of 28% in its future satisfactory selections can be had by use of the proposed test battery, provided only A candidates are selected. In order to enjoy this advantage, however, 100 candidates may have to be screened to get 20. As predictive efficiency diminishes by accepting candidates in lower cut-off score groups, fewer persons need be screened to get the required level of performance. If 50% of present supervisors are considered satisfactory, and management is willing to accept a smaller improvement in selection, say to 80% by using a cut-off score of 68, 70 persons will be located successfully for every 100 screened. By going to the E group, persons cannot be said to be test selected, even though they were given the test since if this group were accepted, all persons taking the test would be accepted, and the test scores would, in fact, be disregarded. In the present study, evidence from the criterion data

suggested roughly a 70% satisfaction with supervisors, so that by inspection of Table XV, the bases examined can expect improvement indicated on the line designated .70. By using only A candidates, 91% of future supervisors can be obtained who will be successful, an increase of 21% or an improvement over present methods of 30%. Table XV can be very useful in explaining to operating officials the value of using the test and of giving preference to persons with higher scores for supervisory vacancies, both realistic problems in using test results in a going organization. It must be cautioned that the improvement in selection efficiency gained by use of the test battery is based on the assumption that all present selection methods will continue to operate as they have been, and that the test will operate as supplementary to the present process.

Expectancy charts. Another method of presenting test efficiency data to management is by way of the expectancy chart which indicates " . . . the probabilities (chances in 100) of various degrees of job success for persons receiving various scores on a particular test."⁵ Table XVI was drawn up on the basis of data available for blue-collar employees at Tinker Air Force Base.

⁵ Dorothy C. Adkins et al., Construction and Analysis of Achievement Tests, (Washington, DC: U.S. Government Printing Office, 1947), p. 163.

TABLE XVI

EXPECTANCY CHART FOR PROPOSED BATTERY
 (Based on blue-collar employees at Tinker AFB)
 (N = 59)

CRITERION	TEST SCORE	
	67 and below	68 and above
Satisfactory or better	17%	62%
Less than satisfactory	14%	7%

Table XVI indicates the accuracy of the test in predicting the job success. In this case the test predicted level of performance correctly 76% of the time, 62% for satisfactory employees and 14% for unsatisfactory; and it erred in only 24% of the cases. From another point of view, if management had used the test in its original selection of present supervisors, 14% of the present workforce or two-thirds of all presently employed, less-than-satisfactory supervisors would have been rejected. The savings to management of the test is made obvious, especially in view of the far-reaching effect on production and employee morale of a supervisory employee.

This same technique of presentation is perhaps even more provocative when only extreme groups are used. Table XVII is based only on the highly successful and outstanding group as compared with the less than satisfactory. Numbers are smaller here so some caution is in order in interpretation.

TABLE XVII

EXPECTANCY CHART FOR PROPOSED TEST BATTERY BASED ON EXTREME GROUPS ONLY (N = 24)

CRITERION	TEST SCORE	
	67 and below	68 and above
Highly Successful. Ratings 7, 8 and 9	4%	46%
Less than Satisfactory, Ratings 1, 2 and 3	33%	17%

For the extreme groups, therefore, it is evident that the test predicted job success accurately 19 out of 24 times times and was in error only 5 of the 24.

Question may be raised concerning the reason for error in prediction in a presumably valid test battery. Of course the test battery is not completely valid. It predicts the criterion with a corrected correlation coefficient of .49 which means that 76% of the variance in the criterion is not accounted for by the test. There are obviously characteristics associated with supervisory success apart from technical competence which are in the evaluations obtained as criteria and which are not measured by the test battery. The predictive efficiency of the battery is not misleading, however, since the data for computing it are based on a sample which had been carefully selected for their positions on interviews, vouchers and observation of personality characteristics, any or all of which may have measured a part of the total variance in the criterion. This gives additional substance to the suggestion that the test battery not be used alone, but only in conjunction with all other screening data available on candidates for supervisory jobs.

Employees. The experience of the Navy Department's use of written tests in selection of blue-collar supervisors has been found gratifying. Employee groups apparently prefer the objectivity of the test

score as a selection criterion over the more subjective considerations frequently used. Pre-clearance with civilian advisory councils where they exist and complete information about the purpose of the test and the exact way in which it will be used may be expected to bring employee confidence to the program. Most federal employees are already familiar with the Civil Service Commission's use of written tests in initial selection, so that this may be considered an extension of that program. When employees fail the test, care will be exercised in explaining the meaning of this fact in terms of their promotability and their future opportunities to qualify through retesting.

VI. SECURITY OF TEST MATERIALS

It is essential that neither the test items nor the scoring keys become common knowledge since the effectiveness of the test battery in discriminating between good and poor potential supervisors becomes completely lost under such circumstances. To prevent this occurrence, the following steps are being taken:

- (1) Alternate forms of the test battery are being readied for release. Sufficient data are available from the research already con-

ducted to develop at least one alternate form. There is naturally less concern over compromise of the "personality" items than of the aptitude items, since knowledge of these items does not provide the respondent with an opportunity to learn the "correct" or keyed answers from other sources. The cost of "personality" item development is nonetheless prohibitive. An alternate form with new aptitude but constant "personality" items is therefore both feasible and sufficient.

- (2) Test booklets and scoring keys, under the regulations, must be kept under lock and key and are to be handled only by authorized persons. Booklets are numbered so that accounting for each copy may be controlled. One person at each installation, designated as the Test Control Officer, is charged with responsibility for assuring security of the test materials.

VII. PROVIDING FOR ECONOMY OF OPERATION

It is the responsibility of a research worker

in an administrative setting to suggest methods for applying his results with maximum economy. All possible efforts were made to meet this objective.

- (1) The tests are group administered, requiring only one test administrator to handle a group of from 30 to 50.
- (2) Test booklets are non-expendable, requiring only a check after each use to remove markings which may have been made on them. Questions are answered on 124 answer sheets which are inexpensive. Where machine scoring is possible, these answer sheets filled in with electrographic pencils may be machine scored.
- (3) The battery has been reduced to the barest minimum of test material which will predict the criterion of supervisory success. There is a minimum of test overlap; consequently a minimum of duplicate measurement and a minimum of testing time.
- (4) By applying unit weighting rather than beta or integral weights, test scores may be compared directly with tables of norms without the additional required step of conversion.

- (5) The directions for administration and scoring are written for the comprehension and use of clerical personnel obviating the need for expensive professional personnel in this routine operation.

VIII. PRACTICAL PROBLEMS OF RESEARCH IN AN ADMINISTRATIVE SETTING

Because administrative considerations impinge on almost every phase of a research study conducted in an administrative setting, it may be of interest to summarize the manner in which those considerations affected this research study.

- (1) Little original item development was possible. Tests in the main were borrowed from other agencies.
- (2) Only two Air Force Bases could be included in the sample, thus affecting representativeness of installation sampling.
- (3) Only 500 supervisors could be included in the sample, thus affecting representativeness of supervisor sampling.
- (4) Only four hours of testing time could be permitted for each supervisor, affecting test selection and need for administering

certain tests at only one base or another.

- (5) IBM machine facilities were not available for a portion of the analysis affecting need for combining subordinate and superior criteria.

In spite of these limitations, which are in no way critical, administrative support was excellent throughout.

CHAPTER VIII

SUMMARY AND CONCLUSIONS

I. SUMMARY

Purpose of study. This study was designed and conducted to investigate the possibility of developing valid selection techniques of an objective nature for selection of civilian supervisors within the Air Force. It was based on the hypothesis that a series of behavioral components exist which can be operationally defined and perceptually and verbally described as supervisory ability, and that this ability can be measured by appropriate scales.

Review of the literature. An intensive survey of the literature in the field of supervisory selection revealed that the problem had been given little attention prior to 1941, but that the many studies accomplished since have suffered from size of sample, faulty criterion data or criterion data collection, inadequate definition of the population sampled and improper methodological approach. Several worthwhile predictors were suggested for inclusion in an experimental battery.

Definition of supervisor. After finding numerous definitions and descriptions of supervisors and their jobs, it seemed that the most uniform and stable approach to defining supervisors for this study would be in such terms as number of persons supervised, occupations supervised and the like. To be included in this study, supervisors must have supervised persons engaged in work in which they (the supervisors) were technically proficient and must have supervised at least three such employees; must have met certain supervision level criteria; and must have been supervisors for at least six months.

The criterion. Considerable care was exercised in collection of criterion data to assure the best possible standard against which to evaluate the experimental test battery. Primarily, sufficient training was given raters to permit them to rate properly; they were both willing and able to make ratings; and they were able to rate on supervisory ability apart from technical competence. Ratings were collected from both subordinates and superiors of supervisors in the sample. These ratings were later combined and found to have a reliability coefficient, corrected by the Spearman-Brown Prophecy Formula of .60. Comparison of the ob-

tained distributions of ratings of both superiors and subordinates with theoretical frequencies revealed a most satisfactory distribution and range.

The sample. The experimental sample included 469 employed supervisors meeting the aforementioned criteria from Hill and Tinker Air Force Bases. In spite of geographical differences in the bases, the mean age and educational levels were not significantly different from each other.

The experimental battery. The tests selected for trial use included a Supervisory Judgment test, three parts of a Personal Preference Inventory, the Rosenzweig Picture Problems, Supervisory Problems, Pattern Matching, Number Facility, a Biographical Information Blank, the Street-Gestalt Test and an Interest Inventory. The "personality" component of the battery was administered uniformly at both bases; the remaining tests were used at only one or the other base to permit the obtaining of maximum information in the allotted four-hour testing period.

Data analysis. To provide for samples to be used as checks on the results obtained, the total sample from both bases was divided randomly into three groups:

one for keying the "personality" tests; one for checking the key and developing a battery; and the last for checking the battery. In keying the tests, both an empirical and an "unconventional" key were used, with two standards applied to each. By selecting only the most predictive sub-tests on this item analysis basis, an r of .32 was obtained with the criterion on an independent sample. Using multiple correlation techniques during the analysis proved unsatisfactory in terms of ability of the resultant weights to hold up when applied on an independent sample. Using rational test selection methods, however, and unit weighting, the most predictive tests did hold up on an independent sample. Adding the most predictive aptitude tests to the "personality" battery, namely, the Supervisory Judgment and Street-Gestalt tests, the correlation with the criterion was raised to .38, which when corrected for attenuation, became .49. This was based only on analysis of blue-collar supervisors.

Problems in use of the battery. To assure standardized administration, scoring and interpretation of test results, specific instructions comprehensible to persons without psychological training were prepared.

Determinations, except concerning security of test materials, are to be made by each Air Force base or Command as to how best to use this test battery in the existing personnel program. Efforts were made to present the validation results in terms most meaningful to management so that they would want to use the test results in making selections and realize the benefits to be derived therefrom. The steps taken to assure economy of operation in using the test were described.

II. CONCLUSIONS

Evidence derived from this study gives credence to the existence of a series of behavioral components known as supervisory ability and subject to operational definition and perceptual and verbal description. This is attested to by the ability of raters to agree about the success of supervisors in the sample in terms of supervisory performance apart from technical competence. That this ability can be measured by appropriate scales is likewise indicated by the ability of the tests finally selected to predict the raters' evaluations. The premise that the resultant multiple correlation could have occurred from chance sampling is discarded on the basis of the fact that the correlation obtained is significant at the 1% level of confidence. And finally

the premise that the final multiple correlation would be high enough for practical predictive efficiency has been demonstrated in the predictive efficiency tables in Chapter VII.

On the other hand, it may not be concluded that the test battery originally selected represents the best possible combination of all existing tests, but rather the best combination of those known to the author at the time. Thus it is entirely conceivable that an experimental battery comprising different tests might have predicted the criterion to a more substantial degree.

The research results derived from the study conducted at two Air Force bases may be generalized to all Air Force Bases, because of sampling considerations going into selection of bases and the populations within the bases, as well as because the concept of supervisory success is relatively standard at all Bases. The final battery is not recommended for application outside the Air Force, however, without check validation. A different pattern of successful supervision might very well require a different set of abilities and consequently a different set of predictors. The methodology and results of this study may, nevertheless, be of value to

other organizations as a point of departure for studies of their own.

III. SUGGESTIONS FOR FURTHER RESEARCH

This research study has been designed to assure reliability of results by attempting to secure adequate sample size, defining objectively the group to be sampled, exercising care in collection of criterion data, and rigorously designing the statistical analysis to provide independent checks at various stages. While the results obtained are somewhat better than many reported in earlier studies, success in terms of validity is still only modest, since it must be recognized that a validity coefficient of .49 accounts for only 24% of the variance in the criterion. This may arise from several causes.

In the first place, personality components are stressed in measuring supervisory success and basic research in personality development is in relatively unadvanced stages. Until more is known about the structure of the complex personality area, its measurement must be equally retarded. Basic research such as that concerned with the authoritarian personality is necessary if measurement is to be highly predictive and if the

"shotgun" approach in the hope that something will work is to be eliminated.

Further than that, relatively little is known about the structure of organizations and the role of the supervisor as viewed from above and below, so that assumptions regarding the constancy of job demands may not in fact hold in every instance. An additional problem arises from the fact that evaluations of success as leaders, particularly from the superior level, may be based on a desire, unconscious or otherwise, to perpetuate in positions of responsibility, the qualities one recognizes in himself so that evaluations of success become highly personal and vary from organization to organization. Thus what may be considered highly related to success by one rater becomes a liability to another. In such circumstances, prediction of a diverse criterion becomes extremely difficult. For example, if a sense of humor were a requirement for success in one case, and a sign of failure in another, a sense of humor test would not predict the criterion viewed in this bipolar fashion. Situational requirements and individual standards, therefore, could well bear intensive investigation.

From the employee's point of view, insufficient is known regarding his concept of supervision or its

affect on his behavior. Do employees sense the supervisor as management's representative or do they produce more favorably under completely employee-oriented supervision and does this also provide a conflicting bipolar basis for evaluation?

From the results of this research study, other areas of research which may be fruitfully attacked suggest themselves. The possibility of analyzing separately the superior and subordinate criteria may lead to better prediction if differences are masked by their combination. The extent to which they disagree makes this possibility worth investigating. More work could be accomplished in the perception of the supervisory job by superiors and subordinates as a basis both for test development and criterion development. Group productivity and sociometric ratings may yield a better rounded picture of supervisory success than the superior and subordinate ratings used alone and in combination lead to higher estimates of validity.

The logic behind occupational differences in supervisory success based on the different needs of blue-collar, clerical and professional employees suggests this as another fruitful area of research.

The importance of the supervisor to government and industry merits intensive study for improved selection.

BIBLIOGRAPHY

A. BOOKS

- Adkins, Dorothy C. et al., Construction and Analysis of Achievement Tests, Washington, D. C.: U. S. Government Printing Office, 1947. 292 pp.
- Adorno, T. W. et al., The Authoritarian Personality. New York: Harper and Brothers, 1950. 990 pp.
- Brown, J. F., Psychology and the Social Order. New York: Rinehart and Company, 1936. 529 pp.
- Cohen, Morris R. and Ernest Nagel, An Introduction to Logic and Scientific Method. New York: Harcourt, Brace and Company, Inc., 1934. 467 pp.
- Deming, William Edwards, Some Theory of Sampling. New York: John Wiley and Sons, Inc., 1950. 602 pp.
- Dennis, Wayne, Carroll L. Shurtle et al., Current Trends in Industrial Psychology. Pittsburgh: University of Pittsburgh Press, 1949. 198 pp.
- Dunn, Paul C., The Selection and Training of Railroad Supervisors. Boston: Massachusetts Institute of Technology, 1942. 58 pp.
- Freeman, Frank S., Theory and Practice of Psychological Testing. New York: Henry Holt and Company, 1949. 518 pp.
- Freeman, Graydon L. and E. K. Taylor, How to Pick Leaders. New York: Funk and Wagnall Company, 1950. 226 pp.
- Fryer, Douglas and Edwin R. Henry, Editors, Handbook of Applied Psychology. New York: Rinehart and Company, 1950. 842 pp.
- Garrett, Henry E., Statistics in Psychology and Education. New York: Longmans, Green and Company, 1938. 493 pp.
- Gowin, Enoch B., Selection and Training of the Business Executive. New York: The MacMillan Company, 1918. 225 pp.

- Guetzkow, Harold, Editor, Groups, Leadership and Men. Pittsburgh, Pennsylvania: Carnegie Press, 1951. 293 pp.
- Guilford, Joy P., Psychometric Methods. New York: McGraw Hill Book Company, Inc., 1936. 566 pp.
- Gullicksen, Harold, Theory of Mental Tests. New York: John Wiley and Sons, Inc., 1950. 486 pp.
- Hemphill, John K., Situational Factors in Leadership. Columbus, Ohio: Ohio State University, 1949. 136 pp.
- Hoslett, Schuyler D., Human Factors in Management. New York: Harper and Brothers, 1951. 327 pp.
- Johnson, Palmer O., Statistical Methods in Research. New York: Prentice-Hall, Inc., 1949. 377 pp.
- Lawshe, Charles H., Jr., Principles of Personnel Testing. New York: McGraw Hill Book Company, 1948. 227 pp.
- Lindquist, E. F., Editor, Educational Measurement. Washington, D. C.: American Council on Education, 1951. 819 pp.
- Mosher, William E., J. Donald Kingsley and Glenn O. Stahl, Public Personnel Administration. New York: Harper and Brothers, 1950. 652 pp.
- OSS Assessment Staff, The Assessment of Men. New York: Rinehart and Company, 1948. 541 pp.
- O'Connor, Johnson, Psychometrics. Cambridge, Massachusetts: Harvard University Press, 1934. 292 pp.
- Pfiffner, John M., The Supervision of Personnel. New York: Prentice-Hall, Inc., 1951. 454 pp.
- Preston, Harley O., The Development of a Procedure for Evaluating Officers in the U. S. Air Force. Pittsburgh, Pennsylvania: American Institute for Research, 1948. 67 pp.
- Riegel, John W., The Selection and Development of Prospective Foremen. Ann Arbor, Michigan: University of Michigan, Bureau of Industrial Relations, 1941. 69 pp.

- Roethlisberger, F. J., Management and Morale. Cambridge, Massachusetts: Harvard University Press, 1946. 194 pp.
- Sanford, Fillmore H., Authoritarianism and Leadership. Philadelphia, Pennsylvania: Stephenson Brothers, 1950. 189 pp.
- Scott, Ellis L., Perception of Organization and Leadership Behavior. Columbus, Ohio: Ohio State University Research Foundation, 1952. 109 pp.
- Stead, Walter S., Carroll L. Shartle and associates, Occupational Counseling Techniques. New York: American Book Company, 1940. 273 pp.
- Stone, Calvin P. and Donald W. Taylor, Annual Review of Psychology. Stanford, California: Annual Reviews, Inc., 1952. 462 pp.
- Thorndike, Robert L., Personnel Selection. New York: John Wiley and Sons, Inc., 1949. 358 pp.
- Thurstone, Louis L., A Factorial Study of Perception. Chicago, Illinois: University of Chicago Press, 1944. 148 pp.
- Tiffin, Joseph, Industrial Psychology. New York: Prentice-Hall, Inc., 1952. 386 pp.

B. PERIODICAL ARTICLES

- Achard, F. H. and Florence H. Clarke, "You Can Measure the Success of an Executive," Personnel, 21: 353-373, May, 1945.
- Beckham, R. O. and Michael Levine, "Selecting Executives," Personnel Journal, 8:415-420, April, 1930.
- Bell, Graham B. and Robert L. French, "Consistency of Individual Leadership Position in Small Groups of Varying Membership," Journal of Abnormal and Social Psychology, 45:764-767, October, 1950.
- Bills, Marion A., "Predicting Managerial Success," Journal of Personnel Research, 4:46-51, June, 1925.

- Bingham, Walter V. and W. T. Davis, "Intelligence Test Scores and Business Success," Journal of Applied Psychology, 8:1-22, March, 1924.
- Blake, Mainwright D. and Arthur E. Barriman, "The Selection and Training of Executives," Journal of Social Psychology, 29:29-33, February, 1949.
- Blankenship, Albert B., "Methods and Problems in Measuring Administrative Traits," Public Personnel Quarterly, 2:69-72, Spring, 1941.
- Browne, C. G., "Study of Executive Leadership in Business: I. The K, A and D Scales," Journal of Applied Psychology, 33:521-526, December, 1949.
- Browne, C. G., "Study of Executive Leadership in Business: II. Goal and Achievement Index," Journal of Applied Psychology, 34:82-87, April, 1950.
- Browning, Rufus C., Kenneth A. Hammond and Frederic T. Wenger, "Self-Selection of Personnel," Public Personnel Review, 12:9-12, January, 1951.
- Carter, Launor et al., "The Relation of Categorization and Ratings in the Observation of Group Behavior," Human Relations, 4:239-254, Number 3, 1951.
- Castle, P. W. G. and F. I. de la P. Garforth, "Selection, Training and Status of Supervisors: I. Selection," Occupational Psychology, 25:109-123, April, 1951.
- Colby, Archie N. and Joseph Tiffin, "The Reading Ability of Industrial Supervisors," Personnel, 27:147-153, September, 1950.
- Comrey, M. L. et al., "Factors Influencing Organizational Effectiveness: I. The U. S. Forest Survey," Personnel Psychology, 5:307-328, Winter, 1952.
- Drury, Lynn B., "Selecting Employees for Advancement," Personnel Journal, 20:166-171, November, 1951.
- Eaton, Joseph W., "Experiments in Testing for Leadership," American Journal of Sociology, 52:523-535, May, 1947.

- Eisenberg, William J., "Qualities Essential for Supervisors," Personnel Journal, 27:251-257, December, 1948.
- Swart, E. E., A. E. Seashore and Joseph Tiffin, "A Factor Analysis of an Industrial Merit Rating Scale," Journal of Applied Psychology, 25:481-486, October, 1941.
- File, Quentin A., "The Measurement of Supervisory Qualities in Industry," Journal of Applied Psychology, 29:323-328, October, 1945.
- File, Quentin A. and H. H. Remmers, "Studies in Supervisory Evaluation," Journal of Applied Psychology, 30:421-425, October, 1945.
- Fraser, John Munro, "An Experiment with Group Methods in the Selection of Trainees for Senior Management Positions," Occupational Psychology, 20:63-67, January, 1946.
- _____, "The Group Method of Selecting Executives," Personnel, 25:50-52, July, 1949.
- Gardner, Burleigh, "What Makes Successful and Unsuccessful Executives," Advanced Management, 13: 116-125, September, 1948.
- Goode, Cecil E., "Significant Research in Leadership," Personnel, 27:342-350, March, 1951.
- Guilford, J. P. and Andrew L. Comrey, "Prediction of Proficiency of Administrative Personnel from Personal-History Data," Journal of Educational and Psychological Measurement, 8:281-296, Autumn, 1948.
- Hanawalt, Nelson G. and Helen M. Richardson, "Leadership as Related to the Bernreuter Personality Measures: IV. An Item Analysis of Responses of Adult Leaders and Non-Leaders," Journal of Applied Psychology, 28:397-411, October, 1944.
- Harrell, Willard, "Testing Cotton Mill Supervisors," Journal of Applied Psychology, 24:31-35, February, 1940.

- Hersey, Rexford B., "Self-Analysis Quiz for Supervisors and Executives," Personnel, 24:454-475, May, 1948.
- Howell, Charles E., "Measurement for Leadership," Sociometry, 5:163-168, May, 1942.
- Humm, Doncaster G. and J. W. Humm, "The Humm-Wadsworth Temperament Scale," Personnel Journal, 12: 314-323, April, 1934.
- Jensen, M. B. and J. B. Motter, "The Validity of the Multiple-Choice Rorschach Test," Psychological Bulletin, 42:182-185, March, 1945.
- Jones, Omer R. and Karl O. Smith, "Measurement of Supervisory Ability," Journal of Applied Psychology, 35:146-150, June, 1951.
- Jurgensen, Clifford W., "Overall Job Success as a Basis for Employee Ratings," Journal of Applied Psychology, 34:333-337, October, 1950.
- Knauff, E. C., "Vocational Interests and Managerial Success," Journal of Applied Psychology, 35:160-162, June, 1951.
- Mandell, Milton M., "Selection of Blue-Collar and White-Collar Supervisors," Personnel, 25:321-327, March, 1948.
- _____, "Supervisors' Attitudes and Job Performance," Personnel, 26:182-183, November, 1949.
- _____, "Testing for Administrative and Supervisory Positions," Journal of Educational and Psychological Measurement, 5:217-228, Autumn, 1945.
- _____, "The Administrative Judgment Test," Journal of Applied Psychology, 34:145-147, June, 1950.
- _____, "The Selection and Promotion of a Personnel Staff," Personnel, 25:125-127, September, 1948.

- Mitchell, J. J., "An Experiment in the Selection of Sales Managers," Occupational Psychology, 12:308-318, Autumn, 1938.
- Moore, Herbert, "Supervision: I. Selection," Personnel Journal, 20:353-356, April, 1942.
- Page, David P., "Measurement and Prediction of Leadership," American Journal of Sociology, 41:31-43, July, 1935.
- Rupe, J. C., "When Workers Hate the Boss," Personnel Psychology, 4:271-289, Autumn, 1951.
- Sartain, A. G., "Relation between Scores on Certain Standard Tests and Supervisory Success in an Aircraft Factory," Journal of Applied Psychology, 30:328-332, October, 1945.
- Schultz, Richard S., "How to Develop Successful Office Supervisors," Personnel Journal, 25:273-281, February, 1947.
- Shartle, Carroll, "Leader Behavior in Jobs," Occupations, 30:164-166, December, 1951.
- Sinaiko, H. Wallace, "The Rosenzweig Picture Frustration Study in the Selection of Department Store Section Managers," Journal of Applied Psychology, 33:36-42, February, 1949.
- Sparks, Charles P., "Limitations of the Bernreuter Personality Inventory," Journal of Applied Psychology, 35:403-406, December, 1951.
- Stockford, Lee, "Selection of Supervisory Personnel," Personnel, 24:186-199, November, 1947.
- Stogdill, Ralph M., "Personal Factors Associated with Leadership: A Survey of the Literature," The Journal of Psychology, 25:35-71, January, 1948.
- Strong, Edward K., Jr., "Differences in Interests among Public Administrators," Journal of Applied Psychology, 31:18-38, February, 1947.

Strong, Edward A., Jr., "Interests of Public Administrators," Public Personnel Review, 6:156-173, July, 1945.

_____, "Interests of Senior and Junior Public Administrators," Journal of Applied Psychology, 30:55-71, February, 1946.

Timpany, Nancy, "Assessment for Foremanship," British Journal of Psychology, 38:23-28, September, 1947.

Uhrbrock, R. S. and Marion W. Richardson, "Item Analysis: The Basis for Constructing a Test for Forecasting Supervisory Ability," Personnel Journal, 12:141-154, July, 1933.

Van Dusen, A. C., "Measuring Leadership Ability," Personnel, 25:67-79, April, 1948.

Wherry, Robert J. and Richard H. Gaylord, "Test Selection with Integral Gross Score Weights," Psychometrika, 11:173-183, September, 1946.

Williams, Stanley B. and Harold J. Leavitt, "Group Opinion as a Prediction of Military Leadership," Journal of Consulting Psychology, 11:283-291, November-December, 1947.

Zeleny, Leslie D., "Objective Selection of Group Leaders," Sociology and Social Research, 24:326-336, March-April, 1939.

C. GOVERNMENT DOCUMENTS

Air Force Manual 40-1, the Civilian Personnel Manual. Washington, D. C.: Headquarters, U. S. Air Force.

Labor Management Relations Act of 1947.

Senate Report Number 2100, 82nd Congress, Supervisory Selection in the Federal Government. Washington, D. C.: U. S. Government Printing Office, 1952. 41 pp.

U. S. Civil Service Commission, Selecting Supervisors. Washington, D. C.: U. S. Government Printing Office, 1951. 30 pp.

APPENDIX A
EVALUATION FORMS

HOW TO USE THE EMPLOYEE EVALUATION BLANK

All of us tend to pass judgments about the people we come in contact with. We might remark to a friend that, "Jack is a nice guy," or, "Bill is a strange fellow." As supervisors, we also have opinions about the ability of our subordinates to supervise their workers. It is an important part of our job. We might feel that, "Jones is about as good a supervisor as you can find," or, "I can't depend on Smith to do anything right." We all know that no matter how good our supervisors are, some of them actually are better than others.

We would like you to evaluate the supervisors who are working for you. But we are going to ask you to make a different sort of judgment than the kind referred to above. What is needed is a judgment which will tell us how much better one man is than another.

Since information is being gathered at other bases and at other units on this base it is necessary that all the judgments be made in the same way. Only then can we make the necessary comparisons among supervisors.

The Employee Evaluation Blank which has been given to you is a device which enables you to rank the supervisors who are now working under your direction. By comparing these ratings with the scores made by these men on the series of tests they will take, we can tell which tests are doing what they are supposed to be doing and how well they are doing it.

You have been given a list of names of men who are scheduled to take the tests and who are working for you as supervisors of other workers. Can you select from the list the man who, in your opinion, has proven himself to be the best supervisor? We know how hard it is to say just what makes a good supervisor—there are as many ideas on this subject as there are supervisors. Although you may not be able to say why you think one man is better than another you may be pretty sure that one man on the list has shown himself to be a better supervisor than the others.

STEP 1. Look at the list of names. You will be asked to choose the man who has shown himself to be the best supervisor—the man you would choose first to fill a new job involving supervision of workers.

Now, which of them would you be least likely to want in a supervisory job? Which one of these men has done the poorest job as a supervisor? We realize that they all may be good men—however it is unlikely that they have demonstrated exactly the same supervisory ability.

Now, write the name of the best supervisor in space 1A and poorest supervisor in space 1B—in the column headed "Supervisor" on the Evaluation Blank. Cross from your list the names just written in spaces 1A and 1B. Again choose the best and poorest supervisor from among those remaining. Write the name of the best supervisor in space 2A and poorest in space 2B. Cross this pair of names from

your list and continue the process of selecting in the same way the best and poorest supervisor from the names remaining on your list. When all the names have been selected, crossed off your list, and entered on the blank, we will be ready for the next step.

STEP 2. Now you will have a chance to indicate how good these men are as individuals. Look at the scales at the left of the Evaluation Blank. There are 9 boxes in each scale. Rate each man by putting an X mark in one of the boxes on the scale next to his name. If you put an X in the last space on the right of the scale--you have said that this man is a nearly perfect supervisor. If you put an X in the first space--on the left of the scale, you have said that this man is a poor supervisor. You can put the X's anywhere on each scale to indicate where this man stands between the two extremes. The closer your X is to the right, the better you think the man is as a supervisor. The closer it is to the left the poorer you think he is as a supervisor. Mark these supervisors, and don't hesitate to ask questions if there is anything you want explained.

STEP 3. Now, in the column headed "Supervisor" where you have already listed the names of your supervisors, draw a circle around the box containing the name of the man whom you consider to be the least effective supervisor, but who is still a satisfactory supervisor. If that clear? Consider just the supervisors whom you consider to be satisfactory, and pick out the one in the satisfactory group that you consider the least effective. Then draw a circle around his name.

THANK YOU FOR YOUR COOPERATION

EMPLOYEE ATTITUDE QUESTIONNAIRE

(Do not sign your name!)

1. What are the two things you like best about your supervisor?

2. What are the two things you like least about your supervisor?

Place a check-mark beside the answer that comes closest to expressing your own opinion about your supervisor.

3. In general, how do you like working for your present supervisor?
 - a. ☐ Not at all
 - b. ☐ Just fair
 - c. ☐ Pretty well
 - d. ☐ Very much

4. Does your supervisor give you enough instruction on how to do your job?
 - a. ☐ He never gives any explanation.
 - b. ☐ He gives some explanation, but not enough.
 - c. ☐ He usually gives enough explanation.
 - d. ☐ He always gives us all we need.

5. Is your supervisor able to answer questions about your work?
 - a. ☐ No, he doesn't know much about my job.
 - b. ☐ He tries but usually doesn't know the answer.
 - c. ☐ He is usually able to answer them.
 - d. ☐ He is almost always able to answer them.

6. When you or other people go to your supervisor with complaints, does he try to improve the situation?
 - a. ☐ He makes no effort.
 - b. ☐ He makes very little effort.
 - c. ☐ He makes some effort.
 - d. ☐ He does all he can.

7. How often do you lose time waiting for instructions or wondering what to do next?
 - a. ☐ Almost all the time
 - b. ☐ Very often
 - c. ☐ Once in a while
 - d. ☐ Almost never

8. How often does your supervisor ask you to do last-minute rush jobs that he could have avoided?
- a. ☐ Most of the time
 - b. ☐ Often
 - c. ☐ Once in a while
 - d. ☐ Never
9. How well do you think your supervisor could do the work you do?
- a. ☐ He couldn't do it at all.
 - b. ☐ He could do it, but not as well as I do.
 - c. ☐ He could do it as well as I do, but no better.
 - d. ☐ He could do it much better than I do.
10. Is your supervisor usually available when you need him?
- a. ☐ Never
 - b. ☐ Once in a while
11. Does your supervisor usually play favorites?
- a. ☐ Yes
 - b. ☐ No
12. About how much does his playing favorites interfere with your work situation?
- a. ☐ It bothers me a great deal.
 - b. ☐ It bothers me some.
 - c. ☐ It doesn't bother me.
 - d. ☐ He doesn't have favorites.
13. How do most of the other people in your work unit like working for your supervisor?
- a. ☐ Not at all
 - b. ☐ Just fair
 - c. ☐ Pretty well
 - d. ☐ Very much
14. How much effort does your supervisor make to train and improve the workers who are too slow or make too many mistakes?
- a. ☐ None
 - b. ☐ Very little
 - c. ☐ Some
 - d. ☐ A great deal
15. Does your supervisor make you feel like putting forth your best effort?
- a. ☐ Never
 - b. ☐ Once in a while
 - c. ☐ Usually
 - d. ☐ Always

16. Does your supervisor treat employees alike regardless of their race, color, or religion?

- a. ☐ Yes
b. ☐ No

17. In general, do you know how you stand with your supervisor?

- a. ☐ Yes
b. ☐ No

18. How important do you consider the following on your job?

	<u>VERY IMPORTANT</u>	<u>IMPORTANT</u>	<u>NOT IMPORTANT</u>
Safe working conditions	☐	☐	☐
Rules about time off	☐	☐	☐
Wages	☐	☐	☐
Pleasant people to work with . .	☐	☐	☐
Supervision	☐	☐	☐
Eating facilities	☐	☐	☐
Medical services	☐	☐	☐
Other things (write in)	☐	☐	☐
☐	☐	☐	☐
☐	☐	☐	☐
☐	☐	☐	☐

19. In general, what characteristics do you think a good supervisor needs?

20. When employees are dissatisfied with their supervisors, what do they usually complain about?

21. Is there anything else you would like to say about the way you are supervised?

22. What is your supervisor's name?

23. How long have you been on your present job?

- a. ☐ Less than six months
b. ☐ Six months or more but less than 1 year
c. ☐ One year or more but less than 2 years
d. ☐ Two years or more

24. Just taking a guess, is your supervisor younger or older than you?
- a. ☐ Younger
 - b. ☐ Older
 - c. ☐ Same age
25. How old are you?
- a. ☐ Under twenty
 - b. ☐ Between twenty and twenty-five
 - c. ☐ Between twenty-six and thirty-five
 - d. ☐ Between thirty-six and forth-five
 - e. ☐ Over forty-six
26. What was the last grade you completed in school?
- a. ☐ Eighth grade or less
 - b. ☐ First year of high school
 - c. ☐ Second year of high school
 - d. ☐ Third year of high school
 - e. ☐ Fourth year of high school
 - f. ☐ More than high school
27. Just one more question about your supervisor. Considering all of his characteristics, how would you rate him as a supervisor? Place a check-mark (✓) in the blank in front of the statement which comes closest to expressing your opinion about how good he is as a supervisor.
- ☐ Unsatisfactory ... about as poor a supervisor as he could possibly be
 - ☐ Unsatisfactory ... but he has a few good points
 - ☐ Barely satisfactory
 - ☐ Satisfactory ... but not quite as good as most supervisors
 - ☐ Satisfactory ... as good as most supervisors
 - ☐ Satisfactory ... a little better than most supervisors
 - ☐ Satisfactory ... he is quite a bit better than most supervisors
 - ☐ Very satisfactory ... one in a hundred
 - ☐ Outstanding ... one in a thousand

THANK YOU FOR YOUR COOPERATION

Check to be sure that you have answered all the questions.

Do not sign your name!

ALL INFORMATION WILL BE TREATED CONFIDENTIALLY.

ProQuest Number: 28783762

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality and completeness of the copy made available to ProQuest.



Distributed by ProQuest LLC (2021).

Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license or other rights statement, as indicated in the copyright statement or in the metadata associated with this work. Unless otherwise specified in the copyright statement or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA