Journal of Criminal Law and Criminology

Volume 14 | Issue 3 Article 5

1924

Aptitude Test for Policemen

Edward M. Martin

AN APTITUDE TEST FOR POLICEMEN

EDWARD M. MARTIN[1]

This study was undertaken by the National Institute of Public Administration with the purpose of determining, by actual experiment, the feasibility of applying to a large body of municipal employees a method of personnel selection which has already demonstrated its advantages in business, industry and in certain other branches of the civil service. Next to education, the largest item in the budget of the average American city is for the police department. To make available a method for devising an entrance examination which will secure better selection of policemen is to make a contribution towards increasing the efficiency of one of the largest city departments. The study was carried out with the co-operation of Mr. Charles P. Messick, secretary and chief examiner of the New Jersey State Civil Service Commission, State Commissioner Edward H. Wright in charge of the Newark district, Captain P. J. Troy of the Fifth Precinct, Newark, New Jersey, and Captain James Meehan of the Newark Police Training School. Further, it was made possible by the technical advice and direction of Dr. Herbert A. Toops of the Institute of Educational Research, Teachers' College, Columbia University. The procedure for evaluating a selective scale, worked out by Dr. Toops, together with formulæ and methods for facilitating computation in the various steps, was followed. We wish gratefully to acknowledge and to express a deep sense of obligation for all such co-operation and direction.

At the present time the selection of policemen by civil service commissions is accomplished through tests of physical and mental fitness. Physical qualifications are determined in terms of standards derived from experience, usually adopted as part of the regulations of the commission. Reliance for the determination of mental qualifications is placed by most commissions in the so-called academic type of examination. While this form is of value as a method of determining one's knowledge of a limited range of specific facts and the ability to express one's self in writing, it has certain defects and disadvantages when used to measure trade aptitude in policemen.

These limitations are revealed when the examination form is analyzed in certain important aspects. Whether such analysis is based on considerations of theory or experimental investigation, a very defi-

[1]National Institute of Public Administration, 261 Broadway, New York City.

nite check on the effectiveness of the form can be had by correlating the ranking of a group in the present entrance examination with a ranking of the same group in a criterion of police ability after a period of service in the department. Such a comparison in the present experimental investigation involving 30 cases yielded a correlation coefficient of —.03 between criterion and an entrance rating which includes credits allowed by law for war service. A comparison which would be fairer to the present test form would be to correlate the criterion with the entrance rating with these war service credits omitted. The comparison in this particular instance yielded a coefficient of —.01.

One aspect of the problem which has undergone extensive analysis is the unreliability of scores in the present examination form. The researches of Starch,[2] Kelley,[3] Inglis,[4] and Ruggles[5] have shown that wide variations in judgment are obtained when a given paper is marked, presumably on the same scale, by two or more examiners. Even when conscious attempts have been made to eliminate these variations, it has been found most difficult to render the scoring of the present examination form objective. The seriousness of this limitation becomes apparent when it is recalled that civil service examination papers are scored by several examiners and that much depends on a proper differentiation between individuals.

It can be said that the present form is often a better measure of handwriting, the physical form of the paper, punctuation and general grammatical form than of individual police aptitude. Men attracted to police work are frequently ill-qualified by previous occupation and training to express themselves in the essay form of answer. Men who otherwise may possess the desired traits may thus be put at an unfair disadvantage. The mental test form of examination, on the other hand, does not eliminate this feature entirely, but does provide a more common basis, since individuals are more alike in their speed of checking, crossing out, underlining, writing single words or short phrases than they are in penmanship or extended written expression.

The range of subject matter covered by examinations may be seen from the accompanying table. This compilation was made from official announcement of examinations or sample forms. Enough cities

[2]Starch, Daniel, "Educational Measurements," p. 9 ff.

[3]Kelley. F. J., "Teachers' Marks"—Teachers College Contributions to Education No. 66.

[4]Inglis, Alexander, "Variability of Judgments in Equalizing Values in Grading," Educational Administration and Supervision, v. 2, pp. 25-30.

[5]Ruggles, Allen M., Study of Unreliability of Raters' Judgments conducted while Service Examiner of the Wisconsin State Civil Service Commission.

are represented to indicate the type of information which it is believed will indicate police ability. In none of the cities, so far as is known, has the validity of the examinations as measures of police aptitude been determined. It is assumed that some relationship exists between the recruit's knowledge of these subjects and his capacity to express that knowledge on paper and his police ability. Such a relationship may exist, but until it has been demonstrated, the basis remains supposition and conjecture, instead of scientific evaluation. A commission's only basis for judging an examination's merits is the judgments of police department officials as to the quality of the men certified. Such estimates are liable to the error inherent in purely subjective personal opinion. As such they are too indirect and too crude a means by which the commission may check the effectiveness of its selective scale when more exact and more effective methods are now available.

TABLE 1

Subject Content of Civil Service Examinations for Policemen in Various Cities

Subject	New Jersey*	New York City	Chicago	St. Paul	Minneapolis	Cleveland	Norfolk, Va.	Missoula, Mont.	Elyria, Ohio	Occurrence
City Information	X	X	X	X	X	X	X	X	X	9
Police Duty, etc...	X	X		X	X†	X†	X	X	X	8
Practical Questions—										
Arithmetic	X	X	X		X	X		X	X	7
Memory Test	X	X		X			X			4
Report Writing				X	X	X				3
Knowledge of Laws and Ordinances ...				X				X		2
Spelling	X		X							2
Geography and Civil Government			X						X	2
Penmanship			X							1
Rules and Regulations.			X							1
General Intelligence ...							X			1

*Examination for cities by state commission.
†Includes duties, terminology, some of laws and ordinances.

The Experimental Group

The co-operation of the Newark police department was enlisted to secure a group of policemen for the experimental evaluation. Being the second largest city in the metropolitan area, it was thought that a group could be secured which would be large enough to satisfy the conditions of the experiment. Also, in view of the fact that the New Jersey State Civil Service Commission was co-operating in the under-

taking as a trial of a new examination method, it was felt that evidence would have considerable value if secured from a representative precinct in the largest city of the state.

The original objective was a group of 100 policemen. If service ratings for each member of the department had been available, it would have been possible to secure such a number. In order to carry out the study it was necessary to build up a criterion and the conditions encountered precluded taking such a large group. Arrangements were made, however, through Chief of Police Michael F. Long, Captain James Meehan of the training school, and Captain Patrick Troy of the Fifth Precinct to secure groups from the school and from the precinct. Two different training school groups were tested. These data could not be included in the experimental study, but are valuable as additional information on the reliability of the tests. These groups, further, served an important function in allaying suspicion and in spreading knowledge as to the general nature and purport of the tests. The result was that the subjects in the experimental groups from the fifth precinct entered the examination with more interest and a better spirit of co-operation than could otherwise have been secured. Forty men were chosen by Captain Troy from his command as subjects for the test group. They were tested in two groups, one on June 20 and the other on June 30, 1922, in the classroom of the training school. The school has quarters in the fifth precinct station house and the accommodations were both convenient and well adapted for examination purposes.

The size of the group was limited by several factors. The forty subjects examined comprised the entire night force of the precinct on regular patrol duty at the time the study was made; it was not expedient for the department for the men from the day force to be included in the study; and a prime consideration in constructing the criterion was that all the subjects should be known intimately to each rating officer. This last condition prevented the inclusion of men from other precincts in the test group. Later it was found advisable to limit the group further because of certain conditions revealed by an examination of all available data. It was considered important to confine the study to the following: patrolmen in active service, men typical of a department, and those who had joined the force under civil service. One subject was found to be a "turnkey" who had done street duty, but had been given a less rigorous assignment with advancing age; an atypical case was found in one man who had been called upon the trial board twelve times in ten years because of neglect of duty; and eight men were found to have joined the force before civil service

became effective. Data of these ten men were, therefore, discarded and the experimental group was reduced to thirty subjects. While it was recognized that conclusions of general application cannot be drawn from so limited a group, it was thought that the data could be used to fill a threefold purpose:

(a) To indicate the feasibility, though on the basis of a limited random sample, of mental tests being used by civil service commissions as selective examinations for policemen;

(b) To serve as material to demonstrate the particular method of research employed in this study;

(c) To make a contribution to the subject of police personnel selection in the hope of leading others to study the problem so that ultimately sufficient data may be had on which to base sound principles of general application.

Establishing the Criterion

A criterion ranking may be secured with relative ease where the individual trade product can be measured quantitatively. But the task is more difficult where the product comprises both quantity and quality. In police work, for example, a patrolman's relative value depends not only on how well he performs routine tasks such as making arrests, box "pulls" or reporting faulty conditions, but also on his attitude towards police work and on the performance of a varied list of functions recognized as police duties. The second group would include such things as: ability to size up a crime situation, influence of his physical presence and personal effectiveness in keeping the district "quiet," acts of service rendered citizens, instruction as to minor law violations, "big brother" influence on the neighborhood children, etc. Routine acts may easily be measured in definite amount; but although other phases of police duty can be measured quantitatively, they are so infrequent in occurrence, often produce results so intangible in character and are so difficult to confirm, that it is only occasionally that they appear on the official records of the department. An account could be kept, for example, of the number of questions answered, instructions given or boys kept "straight" through the friendly intervention of the "cop" on the "beat," but such a tabulation would not be feasible from an administrative point of view. Information is usually available as to the number of arrests, regularity of box calls, or reports made by each patrolman, but these data give too incomplete a picture for them to be considered an accurate measure of police ability. When such official records are lacking, recourse must be had

to the judgment of commanding officers as the criterion of the ability of individual policemen.

The Newark department maintains a record of charges preferred against individual members for neglect of duty or violation of regulations. This record is the nearest approach available in the department to a service rating on members of the force. It is an inadequate criterion of individual ability in that it takes account only of observed neglect of duty and does not consider acts or services of a constructive nature which are performed but do not come to official notice. Also it is only infrequently that charges are preferred against individual policemen so that the measure was not distributed over the group or existed in sufficient quantity to permit its use as a basis for ranking individuals in an experimental group. It was necessary, therefore, to build up a criterion using the judgment of commanding officers as a basis for the scale.

Judgments were secured from four commanding officers—three desk lieutenants and the captain in charge of the precinct. The lieutenants had direct charge of the routine work of the precinct, conducted the daily roll calls and assigned patrolmen to the various beats and received reports from the men on post. They rotated shifts every two weeks so that they had ample opportunity to get an intimate knowledge of the men's qualifications. The captain was continually in contact with the precinct and was in the best possible position to pass an opinion on the value of each man.

Two types of estimates were secured on each of the forty subjects from each of the four rating officers. First, a rating scale was devised for judging the men on four distinct qualifications. Later, a simple ranking in order of ability was secured. The two methods were used as checks for accuracy and as a means of securing an index of reliability for the scale finally adopted.

Moore[6] has summarized under the following headings principles which should be followed in setting up a rating scale:

1. The ability being rated should be analyzed into its component essential abilities or traits and each trait rated independently. It is better to concentrate on a few essential traits rather than to have too many and have the scale break down from its own size and complexity.
2. The traits determined upon must really be different, and as distinct from one another as possible.
3. The rater must be acquainted with the one being rated.
4. The traits must be as sharply defined as possible, so that different raters will rate the same trait.

[6]B. V. Moore, "Personnel Selection of Graduate Engineers," pp. 22-23.

5. The basis of comparison used as a scale should be as concrete and as familiar as possible.

6. Where more than one individual is to be rated in more than one trait, more comparable results are obtained by rating all individuals in one trait before going to the next trait.

7. More reliable ratings result from ratings made independently by more than one person.

The rating scale devised endeavored to apply these maxims and, at the same time, to provide a mode of expression which would secure a proper distribution of judgments. A service rating classification of traits used by the Detroit department was adapted to that purpose. It breaks up the composite "police ability" into the four component traits:

I—Appearance—
- Physique: Athletic or corpulent.
- Neatness: Consider person and dress.
- Bearing: Military attitude and carriage.

II—Intelligence—
- Ability to write a clear and legible report.
- Does he act with good judgment without instructions?
- Does he answer questions intelligently?

III—Discipline—
- Is he punctual?
- Is he respectful to commanding officers?
- Does he obey orders promptly and cheerfully?

IV—Efficiency—
- Does he keep his beat in good condition without arousing ill-feeling among the residents?
- Does he keep his head in an emergency?
- Is he courteous to the public?
- Does he notice violations of ordinances?

To have enumerated an exhaustive list of qualities under each of the four main headings would have served only to confuse the rater; the scale would be too complex and would break down from its own weight. Instead, only outstanding phases were included in order to indicate to the rater the type of qualities to be considered. Thus the general capacity "police ability" is analyzed into four component traits which are not only sharply defined, but are points commonly used by officials in estimating police ability of individual patrolmen.

The rater's estimate is commonly indicated on either a numerical or letter basis. When a percentage scale is used there is an inclination to place everyone in the upper range. It is felt that an injustice is done an individual if he is marked below 70, the usual "passing" grade. If the straight letter basis is used, a grouping results which is too

coarse for accurate statistical purposes. Since neither method insures securing a proper distribution, and results free from preconceived assumptions, it was decided to work out a plan which would as nearly as possible achieve the desired ends. The "human scale" method of estimating individual ability, devised and used by the army for rating officers, was adopted and a stencil device was employed for registering the rater's estimates.

The distinctive feature of the "human scale" is that it utilizes, in systematic and concrete form, the process used more or less consciously in estimating human traits or abilities. Comparisons may be made with abstract standards, but more commonly are made in terms of individuals. Evidence of this practice is found in the expressions, "He is taller than ....," "she is prettier than ....," "he is more intelligent than ....," etc. In other words, the estimate is reached by comparing the person being judged with some one well known to the rater and whose possession of a particular trait is used by him as a standard in such judgments. In short, the values are determined on a human scale. In a casual judgment there are not likely to be more than one or two points on the scale, representing either a single instance or extremes of the trait in question. The scale, however, may be amplified or made precise to any extent desired. Five gradations were used in the present scale and designated as follows: lowest, low, middle, high and highest.

Each rater constructs his own scale by selecting from his acquaintances the individual for each of the five gradations who best represents the particular degree of the trait for which he was chosen. Then, in judging a person, the rater compares him with each of the five men on his own scale. The same scale is used in rating any number of individuals in a given trait. The ratings thus secured can be compared with one another on an equal basis and can be regarded as precise estimates (in terms of the given rater's judgment) of the individual abilities of the several members of a group.

The device employed for registering judgments enabled the raters to disregard entirely any percentage or numerical scale. Each was provided with a cardboard stencil on which are spaces for the construction of a "human scale" for each of the component traits to be rated. The scales on the stencil are arranged to correspond with the listing of the four traits on an individual service rating sheet. The traits, designated I, II, III, IV, are listed down the left hand side of the sheet and on the right hand side, opposite each numeral, is a three-inch horizontal line upon which the rater registers his judgment. The

stencil is just large enough to cover the right half of the rating sheet and has slots or windows three inches long and one-eighth inch wide cut in it so that when lined up with the rating sheet these slots exactly correspond with the four lines on the paper. Each slot is designated with the proper numeral and is spaced off to represent varying gradations of the trait. The points to the extreme left and extreme right of the slot are designated "lowest" and "highest" respectively; the midpoint, "middle"; and the midpoints between the center and the extremes, "low" and "high." A fulcrum is used to indicate each of these points on the scale. Spaces are then provided under each fulcrum for the name of the rater's choice for the particular degree of the trait.

In order to rate an individual, the rater first constructs the four "human scales" on the cardboard stencil. The same individual might appear on more than one scale; the point is emphasized that only the best examples should be used for the respective scales. The stencil is then superimposed on the service rating sheet, care being taken to line it up so that the slots correspond exactly with the three-inch lines on the sheet. With the qualifications of the subject being rated in mind, the rater makes a man-to-man comparison to determine where to place him on the particular human scale. When the range has been narrowed to two men, the rater used his judgment as to just where he falls in the three-fourth inch space between the upper and lower limits of the range. When the exact spot has been determined, the rater inserts a pencil in the slot and marks the distance along the horizontal line on the rating sheet beneath. The process is repeated for each of the other three traits. When the stencil is removed, the subject's degree of qualification in each of the traits is shown by the length of the horizontal line, measured from left to right. The length of line is then evaluated by measuring the distance over a three-inch gauge divided into twenty units. The length of line, as measured in units, thus represents the relative standing of the subject in the trait. The number of points given each trait is indicated along the right hand margin; the separate items for each trait are then totaled and entered at the bottom of the column.

It will be noted that the rater does not need to be concerned with numerical scales of any sort. He has only to make the man-to-man comparison and indicate his judgment in the manner prescribed. The evaluation of points assigned each trait is made by the investigator, and once the rater has registered his judgment, the various traits may be weighted in any manner deemed justified by their relative importance.

In the present study the traits were given the following relative weights:

Trait	Weight
Appearance	20
Intelligence	20
Discipline	20
Efficiency	40
Total Police Ability	100

As a check on the reliability of the first method and to heighten the reliability of the criterion, each of the rating officers was asked to rank the men in the experimental group in order of police ability. Each man's name was typed on a small card. A set of cards, arranged in alphabetical order, was prepared for each officer. The investigator then had a separate interview with each officer in which the need of a second judgment was explained and the procedure to be followed outlined. The cards were first sorted into three piles: average, low and high. The officer was then asked to take each group in turn and to pick out the men as if they were selecting them for a special squad. Choice was made in order of ability, the most desirable—the one in the entire group regarded by the officer as having the most police ability—being selected first, the next best, second, and so on through the entire list of forty. As the cards were picked out, they were handed to the investigator, who arranged them on a table in the order of selection. The cards were lapped so that the name written at the top of each was visible and the whole list of forty names could easily be seen and compared. The officer was then asked to go over the list and to make any changes in the order of the subjects that might occur to him when the list was considered as a whole. The same procedure was followed with each of the four officers co-operating in the experiment.

Two types of judgment were thus secured on each man. There were, therefore, in effect eight sets of estimates of individual ability available for building up the criterion. Before combining the individual estimates into one final score, the two types of estimation were analyzed to determine their relative variability and the consensus ratings and rankings were correlated to determine the reliability of the criterion.

It is possible to determine the variability of any method of estimation by computing the degree of unanimity among any number of raters passing judgment on the same group of individuals. If the four raters, for example, were in perfect agreement as to the position of each of the forty men in the group, their estimate of the relative merits of each would be unanimous. There would be no difference of opinion

between them or their judgments would have a deviation of zero from the general consensus of opinion. The relative variability of the two methods may be determined, then, by obtaining the amount the raters' individual rankings deviate for each method from the consensus ratings and rankings. This computation was made, the individual total scores in the rating having first been turned into serial rank scores so as to be made comparable to the serial ranking obtained by the second method. It showed that the rating method had greater variability than the simple ranking procedure by a ratio of approximately 4 to 3, or 31 per cent. This difference was taken into account in combining the two methods to obtain the criterion by weighting the gross serial rank scores inversely according to the amount of deviation of all the subjects from their consensus ratings and rankings. Each method thus is weighted approximately the same since the standard deviation of the ranks is approximately equal.

The consensus ratings and rankings were correlated and yielded a coefficient of .84, indicating a significant degree of reliability for the criterion thus secured.

Selection of Tests

The problem of what tests to use in the experimental evaluation involves a determination of the requirements of the job in the way of qualifications of its workers. Or, in other words, the discovery of the particular mental qualifications or traits necessary to perform the duties the work requires and the selection of tests that will adequately measure these traits. This determination may be made by a careful analysis of the job in question.

A type of analysis which has had considerable vogue in industry is that which expresses the job requirements in terms of broad human qualities deemed essential in the employee for its proper performance. The job of bookkeeper, for example, would be described as one requiring accuracy, patience, application, neatness, routine temperament, not much initiative or creative ability, concentration, etc. Or the work of an executive would be described as requiring initiative, tact, energy, concentration, etc. Link[7] gives a formidable array of qualities met with in this type of job analysis: dynamic, static, large-dimension or small-dimension worker, industrious, intellectual, volitional, manual, deliberate, impulsive, rapid or slow in mental co-ordination, adaptable, self-centered, roving, settled, loyal, sincere, directive, dependent, re-

[7] H. C. Link, "Employment Psychology," chapter on Job Analysis, pp. 251-269.

sponsible, irresponsible, phlegmatic, live wire, slow but steady, nervously quick, and so on.

In the same way the job of policeman might be said to demand men of such qualities as these (essential physical qualities being assumed): loyal, adaptable, responsible, firm but courteous, intelligent, energetic, sympathetic, incorruptible, honest, routine temperament, etc. Although certain policemen may have these qualities, such a list is not necessarily confined to police work. Several objections, summarized by Dr. Link, may be offered to this type of analysis:

1 It is not job analysis at all, but a thinly disseminated character analysis.

2. The qualities are so general and vague that they mean very little when tied up with a particular job. They are detached and theoretical. A person may have a given quality in one type of work and not in another.

3. They cannot be actually measured except in terms of some concrete job activity. How can loyalty, adaptability, energy, etc., be measured when stated abstractly? No fine distinctions are possible, except in so far as it can be said that a person does or does not have the quality, or to differentiate between good, bad and indifferent.

The method followed in using the mental test technique is in marked contrast to that employed in analyzing jobs in terms of personal qualities. It makes a thoroughgoing analysis of the job and, on the basis of this study, selects a set of tests which seems to involve the same ability as that required by the job. These tests are then tried out on a group of men on the job whose ability is known in order to find out the tests which best measure the specific abilities common to both job and tests. The result obtained by statistical methods is a scientifically accurate measurement of the specific abilities required by the specific job. It is not even necessary to name the abilities; one seeks to approximate them as nearly as possible by the preliminary job analysis, chooses a large number of tests and then relies on the statistical methods to select those tests best adapted to the purpose. The job, then, should be analyzed in terms of concrete job processes or definite functions. The functions may further be analyzed into the mental qualities or measurable traits required in their performance. Tests calculated to measure particular traits are then selected and tried out in the experimental evaluation.

The success of this procedure depends, among other things, on a relatively fixed condition of the job. Such an assumption may be made about police work, for it has become standardized and the duties of a patrolman are fundamentally the same at all times or in all places. An analysis of police work shows that the functions and duties it

involves may be grouped under three headings—I regulative, II investigational, III informational. These may be listed as follows:

I—Regulative duties—
- A—Apprehend criminals.
  - 1—Stop and search suspicious looking characters.
  - 2—Pick up criminals "wanted" from verbal or photographic descriptions.
- B—Enforce laws, ordinances and departmental regulations.
  - 1—Note all such violations.
- C—Occasional duties incidental to patrol.
  - 1—Preserve order at public meetings or assemblages.
  - 2—Parade duty.
  - 3—Maintain fire lines.
  - 4—Traffic at school crossings, etc.
  - 5—First aid relief in emergencies.

II—Investigational duties.
- A—Investigate suspicious looking circumstances and places.
- B—Preliminary investigation of crimes committed on post.

III—Informational—
- A—To citizens.
  - 1—Answering inquiries as to location of streets, hospitals, public institutions; where to address complaints of city service, etc.
- B—To the police and other city departments.
  - 1—Reports of arrests.
  - 2—Reports on crimes committed.
  - 3—Reports on conditions affecting other departments, e. g., broken street lights, broken pavements, etc.
- C—To the courts.
  - 1—Presentation of cases in minor courts.
  - 2—In capacity as witness in criminal cases.

On the basis of this list of duties it is possible to specify certain measurable mental traits which seem essential in their performance. They may be listed as follows:

Memory: Of persons wanted, orders read at roll call, etc.

Observation: Suspicious looking circumstances, conditions on the post, noting circumstances in connection with the commitment of a crime or an accident.

Reasoning analytical judgment: Ability to recognize factors pertinent to a crime situation and solution, ability to piece out a theory of crime from scraps of evidence.

Ability to follow directions: Comprehending and interpreting correctly orders of superior officers.

Ability to organize material for written or verbal report—to desk officer or to a court.

Mental alertness: Ability to size up an emergency quickly and to decide what to do in a given contingency. Police instructions cannot cover every possible contingency, so that the patrolman is forced to judge what course of action is best suited to a situation. Delay in reaching a decision may mean death.
Judgment: Common sense.
Determination: Capacity to persist in a routine assignment.

Several types of tests have been developed to measure the varied aspects of intelligence. No particular test can be said to measure precisely such and such a specific mental trait. The sciences of mental analysis and mental measurement are not yet developed to the point where a specific trait can be singled out and the proper test applied to measure its presence quantitatively. The tests available at the present time tend more to measure traits in combination so that for an experimental evaluation a wide selection of test types should be made. It is combinations of traits, rather than specific abilities, which the practical conditions of a job call forth. To test one particular aspect of intelligence is to test also related traits, for studies of the relation between amounts of desirable single traits show a direct or positive relation between such traits.

Two other factors were considered. One was to include tests covering arithmetic and spelling, subjects used in traditional civil service examinations to determine their value as detectors of police ability. The other was the time limit on each test, since the policemen had been made available for a two-hour period and it was deemed desirable to include a large number of tests covering a wide range of traits in the time available for the purpose. Eleven tests were used and may be listed as follows:

Test 1. Arithmetical fundamentals. Thirty problems in addition, subtraction, multiplication and division selected from the Clifford Woody scales. Time, 5 minutes.

Test 2. Arithmetical reasoning. Twenty problems from the army alpha examination. Only the answers were to be given in spaces provided along the right hand margin. Time, 5 minutes.

Test 3. Spelling. Fifty words, arranged in order of difficulty, common in police parlance. The alternative choice was adopted. Each word was given in two forms—one spelled correctly and the other incorrectly. The correctly spelled words were to be indicated by underlining. Time, 2 minutes.

Test 4. The "opposites" test from the army alpha series. Time, 4 minutes.

Test 5. Number copying. Fifty numbers, ranging in size as follows: the first ten, five digits; the second ten, six digits; the third ten, seven digits, etc., were to be taken in order and copied in numbered spaces on

the reverse side of the sheet. Time, 3 minutes.

Test 6. Common sense test from the army alpha series. Sixteen questions were asked and three answers were given to each. The best answer to each question was to be indicated by a cross placed in a space in front of it. Time, 2 minutes.

Tests 7 and 9. Reading tests devised after the Thorndike reading tests. Test 7 consisted of three paragraphs of crime statistics taken from the annual report for 1921 of the Detroit police commissioner. Fifteen questions were asked based on facts set forth in the report. Time, 12 minutes.

Test 9 was worked out on the same plan, but dealt with a hypothetical crime situation written up in newspaper style. Twenty questions were asked covering specific facts given in the story, deductions from circumstances enumerated and an opinion as to the type of crime described. Time, 15 minutes.

Test 8. A speed test in forming quick decisions designed by Dr. Toops as one of a series of tests to detect stenographic ability. Within the space of a minute as many as possible of 25 questions were to be answered by underlining one of the following: Yes—No—Don't know. The questions were all based on common facts such as "Do you have two hearts?" "Could you lift a bag of feathers weighing 300 pounds?" "Do you take a size 72 night shirt?" Each, however, involved consideration of the correctness of the facts given and deciding which word to underline.

Test 10. The Toops-Pintner Directions Test—one of the standardized tests of ability to carry out specific instructions. It begins with simple directions such as to draw a nose on the figure of a face, and becomes progressively more difficult. Time, 4 minutes.

Test 11. An intelligence test devised originally by Dr. A. S. Otis as part of an examination for policemen and firemen given by the San Francisco civil service commission. It consists of 40 questions, each of which is so phrased that the answer may be indicated by writing either a word, a number or a letter in a bracketed space along the right-hand margin of the sheet. It employs several of the best types that have been found adapted to testing intelligence. The test is constructed on the same plan as the Otis intelligence test. In fact, many questions in the two forms are identical. Time, 20 minutes.

The eleven tests required a total writing time of 73 minutes. It was found that the remainder of the two-hour period was conveniently taken up with explaining the purpose of the experiment, distributing the test forms and pencils, and reading the directions for each of the tests.

Other Possible Variables

A noteworthy feature of the statistical technique employed in this study is the ease with which any factor, common to an entire group and amenable to quantitative measurement, can be analyzed and evaluated. Factors which when considered superficially may not appear to

be pertinent to the inquiry, but which upon statistical analysis may be found to have some contribution to make towards the composite fitness score. Among such factors might be listed the following: social—grade at leaving school, previous occupation weighted according to importance, military experience gauged in length of service; personal—height and weight at appointment to department, age at appointment, number of children, physical measurements[8] such as size of shoes, gloves, collar, hat, etc.; vocational information—length of service in department, grade in police training school, grade received in entrance examinations, merit or demerit marks, rating received in target practice, etc.

It was found possible to secure eleven such factors or variables for the group under discussion. They are enumerated in the following tabulation:

Variable Number	Nature of Variable	Source of Information
12	Grade at leaving school.	Application blank.
13	Height at appointment.	Physical examination record.
14	Weight at appointment.	Physical examination record.
15	Height-weight ratio.	Obtained by dividing height by weight.
16	Age at appointment.	Physical examination record.
17	Police school rating.	Training school records.
18	Length of service in department.	Headquarters records.
19	Civil service entrance rating.	State civil service commission records.
20	Civil service entrance rating minus war service credits.	State civil service commission records.
21	Number of times charges were preferred for violation of departmental regulations.	Headquarters records.
22	Previous occupation.	Application blank.

Explanation should be made of variables 19, 20 and 22. It will be noted that two entrance ratings were obtained from the civil service commission. Variable 19 is the rating which appears on the official records of the commission and is the one used in the appointment of candidates. It includes not only the usual physical and mental ratings, but also war service credits allowed by the New Jersey law (C. 298, P. L. 1920). This law provides that military service credits, ranging from 2½ points to 10 points, shall be added for all veterans with fifteen or more months of war service. Those with a lesser period of service receive a proportionate amount of credit. Wounded or disabled

[8]Physical measurements may have greater significance than is at first apparent. In a recent unpublished paper, Professor H. L. Hollingworth of Columbia University reported a positive correlation between the effect of certain drugs and the quality of organism, indicating a positive relation between physical characteristics and mental ability.

veterans received the maximum number of points. The civil service commission decides the number of credits which shall be allotted in each case.

This provision has no demonstrated relation to the candidate's fitness for police work. Its immediate effect is to give a tremendous advantage to the veteran over the other men in the examination. Eleven men in the test group entered the service with such preference credits. Since it was desired to correlate the ranking of the group by the entrance examination with the criterion, a measure of efficiency on the job, it would be obviously unfair to the civil service commission to include the military credits in the individual ratings. Variable 20, therefore, was worked out by deducting war service credits from each of the individual's official rating and re-ranking the group only on the over-all rating assigned by the civil service examination.

It was considered desirable to work out a method to show the relationship between previous occupation and success in police work. The previous occupations of the men in the group were, therefore, classified arbitrarily according to the degree of mental requirement, as follows:

Weight	Class	Trade
1	Unskilled.	Guard, laborer.
2	Semi-skilled.	Fireman, housewrecker, lather, leather worker, machine operator, molder, railroad man, roofer.
3	Skilled.	Bricklayer, car builder, corset cutter, meter tester, millwright, motorman, riveter.
4	Highly skilled.	Chauffeur, electrician, engraver, machinist, printer.
5	Clerical A	Bartender, freight checker, furniture, house collector, letter carrier.
6	Clerical B	Automobile agent, clerk, salesman.
7	Professional	Newspaperman.

A weight was then assigned each individual in the group on the basis of the above classification.

Complete information was available for each of the eleven variables excepting the 17th and the 19th. In the latter, civil service entrance ratings, ten of the original group of forty men had joined the force before civil service became operative in New Jersey. It was decided, as explained previously, to exclude this group of ten in working up the data since it was desired to include in the final analysis only those subjects on whom complete information in each of the 22 variables was available. To have discarded the civil service entrance ratings as a variable would have rendered impossible a comparison which it was desired to make, namely, the efficiency of the present type of civil service examination in predicting success as policemen. In the

other, training school grade, two subjects went through the training work, but attended when the present captain was on leave. Grades for them were lacking. The average grade for the group was interpolated in these cases on the ground that assigning a value typical of the whole group to a proportionately small part of the whole group will have a negligible influence on the serial order when the cases are ranked according to magnitude.

Examination Groups

The conditions under which the data used in the evaluation were secured were entirely favorable to the success of the experiments. The men were ordered to report for the examination in the same way as any extra duty assignment, such as attendance at police school, parade duty or inspection, might be made. The tests were given from 2 to 4 o'clock on the afternoons of June 20 and 30, 1922. The subjects were from the squads going on duty in the early evening and thus were probably in better physical condition to cope with the examination than if it had been given after their fatiguing tours of duty. The tests were held in the quarters of the police training school, a room in the fifth precinct station house fitted up with school desks and adequate lighting facilities, both natural and artificial.

Although the men were officially ordered to appear for the examination, an effort was made to secure their willing co-operation by explaining the aim of the undertaking, pointing out that it was being done with the approval of the State Civil Service Commission and the sanction of the police department heads, and emphasizing that the mark made in the tests in no way affected their status in the department. That such an appeal was successful is indicated by the fact that no one refused to co-operate and, judging from outward appearances, each man endeavored to do his best in the tests.

During the administration of the tests none but the subjects and the investigator was in the room. Care was taken to minimize distractions. The words "Begin" and "Stop" were used to indicate the time limits for each test. An experience with giving the tests to a training school class prior to the first group of the experimental series led to a slight departure from the standard method of giving the tests. In this first group it was found that the printed directions accompanying each test failed to "get the idea across" with sufficient promptness. This occurred either because of the newness of the undertaking or because of the terse and concise statement of the directions. In order not to risk making a small group smaller, by having any of the men

"soldier" in the tests from failure to understand what was wanted, it was decided to supplement the reading of the printed directions by a short verbal explanation and demonstration of exactly what was to be done in each test. Once the word "Begin" had been given, however, inquiries were discouraged. If any individual persisted, his question was met with the remark, "Use your best judgment."

Evaluation of the Data

In order to derive a composite selective scale which shall correlate most highly with the criterion, one must first compute the necessary intercorrelations between the several variables. The time and labor involved in working out a Pearson r is great, so that it is highly desirable to use any methods which will facilitate the computations to any degree. Particularly is it true of instances such as the present study where twenty-two factors were evaluated in terms of the criterion of ability.

The procedure worked out by Toops[9] was used to compute the required intercorrelations. It reduced the time necessary to make these calculations by using systematic methods which avoid duplication, by the use of gross scores transmuted into numbers so small in magnitude that they may be readily multiplied mentally to obtain the necessary products and the use of a listing adding machine, the printed record from which supplies the required sums for substitution in the correlation formula.

Small numbers may be substituted for large scores in computing correlation coefficients. It has been pointed out by Ayres[10] that the magnitudes of a series of numbers can be reduced without changing their trends. Correlation ascertains the amount of agreement between trends of two series, and these trends are determined not by the size of the numbers, but by the amounts by which the larger numbers in each series are in excess over smaller ones. In the Toops procedure the original gross scores become transmuted scores, none of which is ever larger than 20.

The use of the listing adding machine facilitates the computation in that the printed record displaces all writing ordinarily required in making tally marks, raw scores, etc., found in other methods. It also affords a ready check on the correctness of the work.

In the accompanying tables are presented the intercorrelation co-

[9]H. A. Toops, "Computing Intercorrelation of Tests on the Adding Machine," Journal of Applied Psych., June, 1922, Vol. VI, pp. 172-184.

[10]L. P. Ayres, Journal of Educational Research, June, 1920, pp. 502-504.

efficients which were used in devising the selective scale. Each of the 22 variables was evaluated and the potential contribution of each towards the composite scale was determined. The coefficients of the intercorrelations between the criterion and each of the 22 variables considered are given in Table 3. The coefficients of the intercorrelations between the several variables entering into the selective scale are to be found in Table 4.

TABLE 3

Coefficients of Correlation Between Criterion and Variables

No. of Variable	Name of Variable	Correlation with Criterion
Mental Tests:		
1	Arithmetical Fundamentals	—.249
2	Arithmetical reasoning	—.020
3	Spelling	.245
4	"Opposites"	.151
5	Number copying	.391
6	Common sense	.210
7	Reading tests: statistics	.064
8	Rapid judgment	.247
9	Reading test: crime situation	.307
10	Directions	.313
11	Intelligence	.237
Social, Physical and Personal Characteristics:		
12	Grade at leaving school	—.035
13	Height at appointment	—.240
14	Weight at appointment	.120
15	Weight-Height ratio	.161
16	Age at appointment	—.229
17	Rating in police school	.128
18	Length of police service	.350
19	Civil service entrance rating	—.026
20	Entrance rating minus war service credits	—.010
21	Charges preferred	—.358
22	Previous occupation	—.046

Computation of Multiple Ratio Correlation

The coefficients given in Table 3 are interesting in showing which variables show the closest correspondence with the criterion, but the problem still remains to derive a scale of variables for predicting police ability. We wish to know, further, what is the reliability and validity of such a scale and the relative contribution that each variable has to make to the scale. This latter consideration is dependent on the capacities that the several variables measure.

If each of the tests were of equal importance in detecting police aptitude a combined ranking of the group might be made for all the tests. Or, if it were certain that the capacities involved in answering

TABLE 4

Intercorrelation Coefficients of Component Variables in the Derived Composite

Variable Number	Name of Variable	1	2	4	5	6	9	10	12	13	14	22	Variable Number
	Mental Tests:												
1	Arithmetical fundamentals		.582	.507	.470	.390	.592	.154	-.024	-.171	-.358	.492	1
2	Arithmetical reasoning	.582		.462	.398	.596	.593	.378	-.096	.030	-.125	.474	2
4	"Opposites"	.507	.462		.524	.702	.592	.510	-.029	.055	-.060	.447	4
5	Number copying	.470	.398	.524		.351	.534	.196	-.325	-.095	-.076	.490	5
6	Common sense	.390	.596	.702	.351		.582	.587	-.064	.032	.112	.300	6
9	Reading test: crime situation	.592	.593	.592	.534	.582		.602	-.030	-.232	-.050	.504	9
10	Directions	.154	.378	.510	.196	.587	.602		-.179	-.153	.199	.202	10
	Other Factors:												
12	Grade at leaving school	-.024	-.096	-.029	-.325	-.064	-.030	-.179		.039	-.008	-.135	12
13	Height at appointment	-.171	.030	.055	-.095	.032	-.232	-.153	.039		.370	-.141	13
14	Weight at appointment	-.358	-.125	-.060	-.076	.112	-.050	.199	-.008	.370		-.098	14
22	Previous occupation	.492	.474	.447	.490	.300	.504	.202	-.135	-.141	-.098		22

any one of these tests were different from the capacities involved in answering any other, their relative importance might be based upon the sizes of the coefficients which they separately yielded with the criterion of police ability. But it is possible that different tests involve to some extent the same capacities. The problem, therefore, is to determine the extent of the correlation between any one test and the criterion when everything common to that test and the other tests has been eliminated.

It is also desired to derive a scale of variables by proper selection and scaling of the elements of the individual variables composing the scale, and by weighting the variables in such manner that the weighted composite shall correlate most highly with the criterion.

These determinations may be made by use of a statistical device called the multiple ratio correlation formula, which Dr. Toops has derived from the procedure for computing a true multiple correlation coefficient. This new procedure greatly simplifies the task of deriving such a scale by reducing the process to a series of computations in numbered sequence on a form chart. By following set rules, it is possible to apply statistical methods which heretofore were available only to those skilled in the manipulation of complex mathematical formulæ.

We are indebted to Dr. Toops for the following explanation of the procedure:

"The problem of securing the *maximum* predictive value from a *minimum* number of tests resolves itself into the problem of determining that test, U, of a number of available tests, which will yield by the technique below presented a maximum multiple ratio correlation coefficient when combined at the proper weight with an already existing weighted test composite, C, which is already a maximum. This involves the determination of the correlation of test U with the composite, C; the determination of the weight, β, of test U such that when the deviations in terms of standard deviations in Test U are weighted by that amount the multiple ratio correlation coefficient, r_{IC}', of the *new test composite* at that point shall be a maximum by this method of computation.

"At the outset, that test of the n available tests which correlates highest with the criterion, maximum r_{Iy}, is taken as the "backbone" test, and is thereupon named Test C; whereupon the correlation coefficient r_{IC} is a maximum at the point of *building up* a scale. If the gross scores in Test C are now given a weight of 1.000, there is for

any test, U, a weight, $\beta_{\underline{u}}$, such that when the gross scores in Test U are multiplied by $\beta_{\underline{u}}/\sigma_u$, the multiple correlation coefficient, r_{Ic}', of the two-test scale correlated with the criterion shall be a maximum. But some of the tests, when weighted at their own individual $\beta_{\underline{u}}$ weights, will yield higher multiple correlation coefficients than others. Hence, the ordinary formula for multiple correlation is solved for each of the remaining (n-1) tests in turn:

$$r_{IC} \cdot \sqrt{\frac{r_{IC}^2 + r_{IU}^2 - 2r_{IC} \cdot r_{IU} \cdot r_{UC}}{1 - r_{UC}^2}} \quad (1)$$

That test which yields the maximum multiple correlation coefficient is now called Test U. The weight of Test U with respect to Test C weighted 1.000 is

$$\beta_{\underline{u}} = \frac{r_{IU} - r_{IC} \cdot r_{UC}}{r_{IC} - r_{IU} \cdot r_{UC}} \quad (2)$$

We now consider that theoretically the gross scores on Test U have been added at the proper weight to the gross scores of Test C; our problem is then to find by formula the weight of a new test U′ such that when its gross scores are added to the now existing test composite, C′ (the gross scores of which are considered as now being weighted 1.000) the multiple ratio correlation coefficient shall be a maximum for all the possible remaining (n-2) tests. It is not necessary to actually combine the gross scores and compute the necessary correlation coefficients, $r_{U'C'}$, since a formula obtains the same result:

$$r_{U'C'} = \frac{\Sigma r_{U'y} \cdot W_y}{\sqrt{\Sigma W_y^2 + 2\Sigma r_{xy} \cdot W_x \cdot W_y}} \quad (3)$$

in which, $r_{U'y} \cdot W_y$ is the sum of the single-products of all the correlations of Test U′ that occur *in a column*, U′, each respectively multiplied

by the weight of the test, W_y, of the row for all the rows of the test composite C′ which enter into a double symmetrical intercorrelation table, which is being built up as the test composite is being built up. This test composite at this point consists of Tests C and U, whence the two correlation-products are $r_{U'C}.(1.000)$ and $r_{U'U}.\underset{C}{\beta u}$.

"ΣW_y^2 is the sum of the square of the weights in the test composite at this time, namely, here, $(1.000)^2 + (\underset{C}{\beta u})^2$.

"$\Sigma r_{xy} W_x W_y$ is the sum of all the double-products when each of the *intercorrelations of tests* in the test composite at this time are respectively multiplied by the weight of the row and column in which they are found in an intercorrelation table, namely, here, $r_{uc}.(1.000)\ (\underset{C}{\beta u})$.

"With this equation solved for all the remaining (n-2) variables, resort to formula (1) determines which test will yield the maximum multiple ratio correlation coefficient, r_{IC}". This test, when determined, is called U". The weight of U" in the multiple ratio regression equation is given by the formula

$$\beta' = \sqrt{\Sigma W_y^2 + 2\Sigma r_{xy} W_x W_y}\ \frac{r_{IU'} - r_{IC'} r_{U'C'}}{r_{IC'} - r_{IU'} r_{U'C'}} \tag{4}$$

The quantity under the radical does not enter into formula (2) for the reason that the standard deviation of the original test C is 1.000, when measured in terms of its own standard deviation. When adding on Test U′ the composite C′ has a standard deviation of its own which must be considered, and which this radical expression takes full account of. Equation (4) is the perfectly general expression of the weight at which a new test U′ is added to an already existing test composite C′. By repetitions of the procedure involved in adding Test U′ as above outlined, one may determine in succession the fourth, fifth, sixth tests, and so on. The multiple ratio correlation coefficient at each point is an index of the efficiency of the scale. Soon a point of diminishing returns is reached, where the addition of a test adds but little to the value of the multiple ratio correlation coefficient at that point, and the value available will approach the value which we would receive from the inclusion of the entire *n* tests. At this point the test can be con-

sidered complete. Any two or more of the tests can be used for the scale by always cutting off from the composite any number of the tests which are added last, whereupon that part of the multiple ratio regression will apply which is left after the exclusion of any number of tests at the right hand end of the equation. The general equation for combining the gross scores of any number of tests is:

$$\frac{X_I - M_I}{\sigma_I} = \frac{X_c - M_c}{\sigma_c} + \beta \frac{r_{ic'}}{6} \frac{X_u - M_u}{\sigma_u} + {}' \frac{r_{ic''}}{} \frac{X_{u'} - M_{u'}}{\sigma_{u'}} + {}'' \frac{r_{ic'''}}{} \frac{X_{u''} - M_{u''}}{\sigma_{u''}} + \cdots\cdots \quad (5)$$

The cumulative correlation after the addition of each test is shown above each respective test at which the cumulative correlations become available. The multiple ratio correlation coefficient is not a true multiple correlation coefficient, but is a very close approximation to it."

The method is, in effect, an elimination process in which those variables calling out abilities that are identical with those considered desirable in policemen are revealed and all others are rejected. The efficiency of given tests in measuring various mental traits is gauged by the size of the correlation coefficients. A high correlation between the test and the criterion indicates it to be an efficient instrument. A high correlation between tests indicates that both are measuring to a large extent the same trait or combination of traits. The ideal condition, then, in building up a scale of tests is to have high correlation between tests and criterion, and low intercorrelation between tests. The magnitude of the r_{IC} is the index of the reliability of the scale of variables built up. Since it also determines the accuracy of prediction, the more reliable the scale, the better will it be as a means of predicting success on the job for candidates on the force.

In evaluating the variables in the present study, it was decided to build up a scale of mental tests first, and then to add in the remaining variables obtainable at the time of application to determine what value they might have in such a selective scale. Variable 5, the number copying test, gave the highest correlation with the criterion, .39, and was chosen as the "backbone" for the combination. Successive solutions of the formulæ led to the selection of the following variables and composite fitness score for the selective scale thus constructed:

TABLE 5

The Contribution of Each Variable to the Efficiency Index and the Weighting of Each Variable in the Derived Scale

Variable Number	NAME OF VARIABLE	EFFICIENCY INDEX OF SCALE		WEIGHTING OF COMPONENT VARIABLES			Length of Test in Minutes
		Cumulative Multiple Ratio Coefficient	Increment in Coefficient	Standard Deviation (σ)	Beta	In terms of Gross Scores 30d	
(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
5	Number copying.....	.39	.....	.....	1.000	.....	.....
1	Arithmetical fundamentals.	.63	.24	4.356	—.852	—5.87	5
9	Reading test: crime situation	.69	.06	4.822	.489	3.04	15
2	Arithmetical reasoning.....	.70	.01	3.511	—.223	—1.91	5
10	Directions test.....	.72	.02	3.182	.229	2.16	4
4	"Opposites".....	.72	.00	10.602	—.164	—.46	4
6	Common sense.....	.73	.01	3.438	.139	1.21	2
5*	No. copying (reweighted)...	.74	.01	3.790	.735	5.82	3
	Length of Mental Test:						38
13	Height at appointment.....	.77	.03	1.213	—.239	—5.92	
12	Grade at leaving school....	.78	.01	1.999	.223	3.35	
22	Previous occupation.....	.79	.01	1.498	—.157	—3.15	
14	Weight at appointment....	.80	.01	13.044	—.102	—.24	

*Test 5 was reweighted by subtracting its product components in the numerator and denominator and resolving for r_{1C}'.

It will be seen that an examination of seven tests, requiring 38 minutes to complete, gave an efficiency index for the composite of .74. Then, by taking into account four measures available when recruits join the force, this score is raised to .80, a correlation coefficient of significant magnitude. The fact that r_{1C} was increased from .74 to .80 by taking into account the variables, education, previous occupation and height and weight at appointment is to be interpreted that these factors had a very definite bearing in determining ultimate success as policemen for the thirty men in the experimental group.

In column G of Table 5 are given the weighting of each variable included in the selective scale. These values, five of which are positive and six of which are negative, represent the relative contribution which each test makes towards the total fitness score. Positive credits are allotted for possession of certain desirable traits, while the possession of other qualities is found to have an inverse relation to police ability and is therefore penalized by demerit credits.

In column F are given the Beta weights obtained in the solution of the formula. These, however, are in terms of transmuted scores and are based on the conditions of each variable. The variability of the array of cases in each variable is different, so that to make the weights comparable each must be divided by a measure of the variability of the test it represents. The standard deviation (σ) is such a

measure. To divide each weight by the standard deviation of its variable (shown in column E) is to eliminate the factor of variability, and to make the various test weights comparable. This computation was made for each weight and the result obtained in each case was a four place decimal fraction. To turn these weights into numbers which could be more conveniently handled in weighting gross test scores, each weight was multiplied by 30. The resulting products appear in column G of the table.

Practical Application of Results

In a practical application of the results of this study, candidates for positions on the force would be given a 38-minute examination composed of these tests: 1, 2, 4, 5, 6, 9, 10. They would also be required to submit their qualifications in the following subjects: grade at leaving school, indicated by last full year completed on scale 1-16; their net height and weight at the time of application; and their occupation prior to time of application. Their scores in each of the subjects would then be determined (number of correct answers in each mental test, school grade completed, height in inches, weight in pounds and the grade of previous occupation), and would be modified in accordance with the proper weighting of the factor. This could be done by multiplying each of the eleven scores by the proper weighting, and getting the algebraic total of the individual weighted scores which would give the relative standing of the individual on an established scale. The same result may be achieved, with greater dispatch and less likelihood of error, by constructing a table of fitness scores which will give values for all possible individual test scores.

In devising this table the total range of possible gross scores for each variable would be determined and scales of weights would be computed. In the seven mental tests the range would extend from 0, for total failure, to the number of credits it would be possible to receive in any one test. Test 1 has 30 questions, so that the lowest possible number of credits would be 0 and the highest 30, if one point is allowed each correct reply. Test 5 has 50 numbers, so that the range is from 0 to 50. The range for schooling would be from 0, no formal schooling, to 19, the number of years required normally to achieve a doctorate, the highest academic degree granted; for height, from 5′7″ to 6′6″, the minimum and maximum heights permitted under the present New Jersey regulations; for weight, 144 to 246 pounds, for similar reasons; and for previous occupation, 1 to 20, on the basis of 20 occupational classes grouped according to intellectual requirements. The same gross score scales may be used for the mental tests and schooling.

Special gross score scales were worked out for height and weight since, though both are measured numerically, their intervals and ranges do not coincide.

It will be noted that positive credits are used for both positive and negative weights. It was thought that instances of the inverse relationship presented were so unusual that to assign negative credit would serve to cause confusion in weighting test results. It was decided, therefore, to work out a method in which only plus credits would be used.

The method used provides for inverting the scale of credits for negative weights. Thus, instead of giving the most credit to the best performance and the least to the poorest and then subtracting the score for negatively weighted traits, the process is just reversed: the poorest performance gets the greatest number of credits, and the best, the least. The system of weighting traits is not impaired, for the "best" performance still receives the highest amount of credit.

The practicality of giving negative weights may be questioned from the administrative standpoint. That is, if the system of weighting became known, would not applicants "stall" on those tests in which the amount of performance is in ratio to possession of the aptitude measured? In Tests 1, 2 and 4, for example, zero scores receive the most credit. The answer to such an objection is that the method of weighting need not be advertised and proctors at examinations could uncover individual instances of hypocritical performance. The relative weights of each factor in the selective scale may be computed on a percentage basis. The relative amounts of the weighting could then be published in the examination prospectus, as is now done, without reference to how the credits would be assigned. And the paper of any individual whom the proctors detected in "soldiering" on a test could be forthwith refused further consideration. It may be pointed out, too, that of the eleven factors only three would give the difficulties just mentioned. These are the three tests in the mental examination. Of the other three, height and weight could be determined easily, and an affidavit with references would secure adequate and reliable information as to previous employment.

Checking for Error in Computation

The accuracy of all computation involved in the procedure may be tested by weighting the original scores with proper credits from the fitness table, and then correlating the individual total weighted scores with the criterion. If no errors have been made in the various steps—transmutation of scores, intercorrelation of tests, and multiple ratio

correlation computation—the coefficient of correlation thus obtained will be the same as the multiple ratio composite fitness score of the selective scale worked out.

This check was applied to the series of seven mental tests and then to the entire scale of eleven factors. In the first instance, individual weighted totals of only the mental tests were correlated with the criterion and the individual totals of tests, schooling, previous occupation and height and weight were correlated with the criterion. The coefficients thus derived agreed within .01 with the multiple ratio coefficients obtained for the respective lengths of selective scale. The following tabulation gives the comparison:

Length of Scale	Multiple Ratio Coefficient	Correlation Weighted Totals with Criterion
Six Mental Tests	.74	.73
All Eleven Factors	.80	.81

The discrepancy is so small as to be negligible so far as it is an indication of serious error in the computation. If the rs were given to only one decimal place, the agreement would be perfect.

Advantages of Procedure

When contrasted with the prevailing method of selecting policemen, the procedure outlined herein has several advantages which indicate what can be accomplished by systematic and scientific methods. These advantages may be outlined as follows:

A—Setting Up Examination: On basis of job analysis and experimental evaluation.

B—Examination Form:

1—Range: Question form permits increased range of inquiry.

2—Scoring:

a—Objectivity: Frees grading from personal bias of rater; results in a given paper receiving the same score any time and from any rater.

b—Use of Scoring Key: Arrangment of questions and replies permits use of stencils and scoring keys, which reduces scoring of examinations to routine procedure. Scoring may be by clerks, thus relieving the examiners for more important work.

3—Weighting: The weighting of each component test in the composite scale is accurately determined by statistical methods.

4—Merit of Scale: The validity of the selective combination, as measured on a known scale of value and in terms of an accepted criterion of police ability, is definitely known. Thus, in revising the entrance examination, as occasion may require, it is possible to gauge the amount of improvement over previous combinations.