# MULTIPLE FACTOR ANALYSIS

BY L. L. THURSTONE
*The University of Chicago*

The two-factor problem of Spearman consists in the analysis of a table of intercorrelations for the discovery of some general factor that is common to all of the variables in the table. Spearman differentiates three types of factors, namely, a general factor which is common to all of the variables, group factors which are common to some of the variables but not to all of them, and specific factors that are peculiar to single variables alone. In practice, the Spearman two-factor methods meet with the difficulty that group factors are frequently encountered. The two-factor methods are not applicable to situations that involve group factors except in indirect ways. This is a serious limitation on Spearman's technique since many important psychological problems involve a complex of variables that are known from the nature of the problem to contain group factors. The present multiple factor methods in no way contradict the Spearman two-factor methods which are very ingenious and powerful in the situations to which they apply. The present multiple factor method may be thought of as supplementary to the Spearman two-factor method in that we do not have any restrictions as to the number of general factors or the number of group factors.

It is the purpose of this paper to describe a more generally applicable method of factor analysis which has no restrictions as regards group factors and which does not restrict the number of general factors that are operative in producing the intercorrelations. Our first question concerns the number of general, independent, and uncorrelated factors that are operative in producing a given table of intercorrelations for any number of variables. In our terminology general factors will include what Spearman calls general and group factors.

This question can be answered by methods that will here be described and which are applicable to any table of intercorrelations. We may consider three examples to illustrate the nature of this first problem. If we have a table of intercorrelations for a battery of motor tests it is of considerable psychological interest to know how many independent motor abilities it is necessary to postulate in order to account for the whole table of intercorrelations. If this turns out to be three, then our next task would be to hunt about for the nature of these three motor abilities. If we have a table of intercorrelations of the interests of eighteen professions it would be of considerable importance to know how many independent interest factors it is necessary to postulate in order to account for the whole table of intercorrelations. This refers to the tables published by E. K. Strong. We have applied our methods to his data and we have found that his table of intercorrelations can be accounted for by postulating four general interest factors which turn out to be (1) interest in science, (2) interest in language, (3) interest in people, and (4) interest in business. Again, Professor Moore [1] has prepared a table of intercorrelations of 48 psychotic symptoms on the basis of his work with about four hundred patients with various psychoses. A general factor analysis of the type here discussed would enable us to know how many general factors or mental disease entities it is necessary to postulate in order to account for the whole table of correlations of psychotic symptoms. If this should turn out to be five, for example, then we should be justified to look for five fundamentally different psychoses.

Our next problem is to assign a weight or loading of each of the general factors to each of the variables. For example, in the table of interest-correlations above referred to we should assign four loadings, one for each of the four general factors, to each of the eighteen professions. It then turns out that Engineering, for example, has a high loading of interest in science, a rather low loading of interest in language. The profession of law has just the reverse loadings, namely low

[1] T. V. Moore, The empirical determination of certain syndromes and underlying praecox and manic depressive psychoses, *Amer. J. Psychiat.*, 1930, 9, 719–738.

for science and high for language. The ministry is loaded high for interest in people and in language but low for science. Finally, we should want to be able to assign to each individual subject a quantitative rating in the form of a standard score for each of the general factors or abilities that have been isolated.

Let there be $n$ factors. In this explanation $n$ will be assumed to be three. It can be any number.

Let there be $N$ individuals in a group, all of whom have taken $w$ tests.

Let $a$, $b$, $c$, $d$, etc., represent the tests.

Let the three factors be represented by numerals 1, 2, 3.

$S_a$ = standard score of one individual in test $a$.

$S_b$ = standard score of one individual in test $b$.

$S_c$ = standard score of one individual in test $c$.

$S_a = \frac{X_a - m_a}{\sigma_a}$ = the usual definition of a standard score.

The standard score $S_a$ of an individual in test $a$ depends on (1) his rating in each of the three abilities or factors 1, 2, 3, and (2) the weight or loading of each of these abilities in test $a$. For example, if test $a$ calls for much of ability No. 1 and very little of abilities 2 and 3, and if one subject has a low rating in ability No. 1 and average or high ratings on the other two abilities, then this subject may be expected to do poorly on test $a$. The loadings of the three general factors in each test and for each subject may be represented by the following notation.

Let

$x_1$ = standard score of an individual in ability No. 1.

$x_2$ = standard score of an individual in ability No. 2.

$x_3$ = standard score of an individual in ability No. 3.

and let

$a_1$ = loading of ability No. 1 in test $a$,

$a_2$ = loading of ability No. 2 in test $a$,

$a_3$ = loading of ability No. 3 in test $a$.

Then we shall assume that the standard score of each individual subject is a sum of the products of his standard score in each ability and the loading of the ability in each test. This assumption leads to the following fundamental equations.

$$S_a = a_1x_1 + a_2x_2 + a_3x_3 \qquad (1)$$

and

$$S_b = b_1x_1 + b_2x_2 + b_3x_3.$$

Strictly speaking, we should add to each of these two expressions a term to account for those additional general factors, beyond three, which are here ignored, and also a term to account for the specific factor, peculiar to the particular test. However, our object is to ascertain how many general and independent factors it is necessary to postulate in order to account for a whole table of intercorrelations and we shall therefore intentionally ignore these additional specific factors as well as those minor group factors which may not appreciably affect the correlations.

We want to express the correlation between tests $a$ and $b$ in terms of the standard scores in the three abilities and the loadings of the three abilities in each of the two tests. For this purpose we shall need the product $S_aS_b$. Then

$$S_aS_b = a_1b_1x_1^2 + a_2b_2x_2^2 + a_3b_3x_3^2 + \text{cross products.}$$

The correlation $r_{ab}$ can be expressed simply as

$$r_{ab} = \frac{\Sigma S_a \cdot S_b}{N},$$

because $S_a$ and $S_b$ are both standard scores so that the standard deviations of the given scores are all unity. Then

$$\frac{\Sigma S_aS_b}{N} = a_1b_1\frac{\Sigma x_1^2}{N} + a_2b_2\frac{\Sigma x_2^2}{N} + a_3b_3\frac{\Sigma x_3^2}{N},$$

in which the cross products vanish because $x_1$, $x_2$, $x_3$ are uncorrelated. But

$$\frac{\Sigma x_1^2}{N} = \frac{\Sigma x_2^2}{N} = \frac{\Sigma x_3^2}{N} = 1,$$

since $x_1$, $x_2$, $x_3$ are all standard scores in the three abilities. Therefore

$$r_{ab} = a_1b_1 + a_2b_2 + a_3b_3. \qquad (2)$$

This is one of the fundamental equations. Here the correlation between two tests is expressed in terms of the loadings of the three abilities in the two tests. By analogy we may write equation (2) for any pair of tests, as

$$r_{ac} = a_1c_1 + a_2c_2 + a_3c_3$$

and so on.

Another fundamental equation can be derived as follows:

$$S_a = a_1x_1 + a_2x_2 + a_3x_3.$$

Squaring,

$$(S_a)^2 = a_1^2x_1^2 + a_2^2x_2^2 + a_3^2x_3^2 + \text{cross products.}$$

Summing and dividing by $N$

$$\frac{\Sigma S_a^2}{N} = a_1^2\frac{\Sigma x_1^2}{N} + a_2^2\frac{\Sigma x_2^2}{N} + a_3^2\frac{\Sigma x_3^2}{N}.$$

The cross products vanish because $x_1$, $x_2$, $x_3$ are uncorrelated by definition. But

$$\frac{\Sigma x_1^2}{N} = \sigma_1^2 = 1$$

and

$$\frac{\Sigma S_a^2}{N} = 1 \text{ by definition.}$$

Hence

$$a_1^2 + a_2^2 + a_3^2 = 1, \tag{3}$$

and similarly by analogy for every other test, as

$$b_1^2 + b_2^2 + b_3^2 = 1.$$

Equations (3) come about because $x_1$, $x_2$, $x_3$ and $S_a$, $S_b$ are standard scores.

Still another fundamental equation that we shall need can be derived as follows:

$$\begin{array}{rcccccccc}
r_{aa} &=& a_1^2 &+& a_2^2 &+& a_3^2 &=& 1. \\
r_{ab} &=& a_1b_1 &+& a_2b_2 &+& a_3b_3 & & \\
r_{ac} &=& a_1c_1 &+& a_2c_2 &+& a_3c_3 & & \\
r_{ad} &=& a_1d_1 &+& a_2d_2 &+& a_3d_3 & & \\
\cdot & & \cdot & & \cdot & & \cdot & & \cdot \\
\hline
\Sigma r_{ak} &=& a_1\Sigma k_1 &+& a_2\Sigma k_2 &+& a_3\Sigma k_3 & &
\end{array} \tag{4}$$

In the summation equation $\Sigma k_1$ = sum of all loadings of the first factor in all of the tests of the set. The notation $k$ refers to each of the tests in succession, *i.e.*, $k$ takes values from 1 to $w$ when there are $w$ tests in the series. But this summation equation may be written for each one of the tests as follows

$$\begin{aligned}
\Sigma r_{ak} &= a_1\Sigma k_1 + a_2\Sigma k_2 + a_3\Sigma k_3 \\
\Sigma r_{bk} &= b_1\Sigma k_1 + b_2\Sigma k_2 + b_3\Sigma k_3 \\
\Sigma r_{ck} &= c_1\Sigma k_1 + c_2\Sigma k_2 + c_3\Sigma k_3 \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
\Sigma r_{kk} &= (\Sigma k_1)^2 + (\Sigma k_2)^2 + (\Sigma k_3)^2 \qquad (5)
\end{aligned}$$

in which

$\Sigma r_{kk}$ = sum of all intertest correlations in the whole table. Since the full table is symmetrical it follows that each intertest correlation enters twice in this sum. Also, the full table of intercorrelations includes the correlation of each test with itself and this is here regarded as unity.

$\Sigma k_1$ = sum of all loadings of factor No. 1 in all tests.

$\Sigma k_2$ = sum of all loadings of factor No. 2 in all tests.

$\Sigma k_3$ = sum of all loadings of factor No. 3 in all tests.

Note that equation (3) is the equation of a sphere and that $a_1$, $a_2$, $a_3$ are the three coördinates of a point on the surface of the sphere. The space order of the sphere is equal to the number of postulated general factors. If five factors were postulated we should have a five-dimensional sphere, and so on.

Also note that each test is here really a point on the surface of a sphere. Hence, if there are three factors essentially operative in producing the intercorrelations, then it should be possible to locate each test as a point on the surface of a ball.

Note also that each correlation is the cosine of the central angle subtended by the two points $a$ and $b$. Hence, if three factors are operative, then each test would be represented by a point on the surface of the ball, and the points would be

so allocated that the cosine of each central angle is equal to the intercorrelation between the respective pair of tests. This will be shown in more detail later.

Our problem is to ascertain the coördinates $a_1$, $a_2$, $a_3$ for test $a$, the coördinates $b_1$, $b_2$, $b_3$ for test $b$, and so on for each of the $w$ tests in the whole series. Then, if our determinations are correct, it should be possible to calculate the correlation coefficients by equations (2). These calculated coefficients should agree with the observed correlations within reasonable experimental error.

If we had all of the tests allocated to as many points on the surface of a ball we could not determine the coördinates for any of these points without first deciding where our coördinate axes are to be drawn. The location of these axes is arbitrary and not at all given by the intercorrelations because the latter are merely the cosines of the angular separations between all pairs of points on the surface of the ball.

One simple plan would be to draw the axis $OX$ through one of the tests which might be more or less arbitrarily chosen, such as $a$. Then if this $x$-axis represents the first factor, it follows of course that the $xyz$ coördinates of point $a$ are $(+1, 0, 0)$ and hence that

$$a_1 = +1, \quad a_2 = 0, \quad a_3 = 0.$$

The $y$-axis must of course be at right angles to the $x$-axis but it could be drawn through the origin in any direction in the plane at right angles to the $x$-axis.

We might now revolve the sphere around the $x$-axis until any arbitrarily selected second test $b$ lies in the $xy$-plane. Then it is clear that the $z$-coördinate of point $b$ must be zero so that $b_3$ in equation (3) vanishes. Therefore equation (3) becomes

$$b_1^2 + b_2^2 = 1$$

and

$$a_1 = 1.$$

The correlation between $a$ and $b$ is

$$r_{ab} = a_1b_1 + a_2b_2 + a_3b_3,$$

but since $a_1 = 1$, $a_2 = 0$, $a_3 = 0$, $b_3 = 0$, this reduces to

$$r_{ab} = b_1.$$

But

$$b_1 = \cos \phi$$

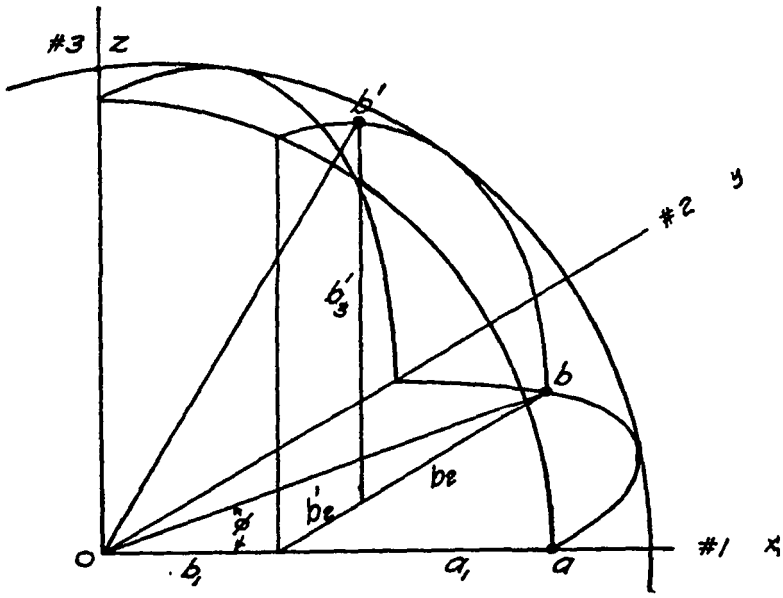


FIG. 1. This diagram shows the relation between the correlation coefficient and the central angle between the two tests or points. Let two points which represent any two tests be designated $a$ and $b'$. Let the $x$-axis, representing the first general factor, pass through the point $a$. Then its coördinates are (1, 0, 0). Therefore $a_1 = 1$ while $a_2$ and $a_3$ are both zero. Revolve the sphere about the $x$-axis so that the point $b'$ is in the horizontal $xy$-plane at $b$. Then the $z$-coördinate of the point $b$ is evidently zero while $b_1^2 + b_2^2 = 1$. Let the angle $aob$ be designated $\phi$. Then, clearly, $\cos \phi = b_1$ since the radius of the sphere is unity. The correlation between tests $a$ and $b$ can be written as follows.

$$r_{ab} = a_1b_1 + a_2b_2 + a_3b_3.$$

Since $a_2$ and $a_3$ are both zero, the second and third terms vanish and since $a_1 = 1$, it follows that $r_{ab} = b_1$. The angle $aob$ is equal to the angle $aob'$. Hence the correlation coefficient is equal to the cosine of the central angle between the points that represent the two tests.

and hence the correlation between tests $a$ and $b$ is the cosine of the central angle between them. Since the sphere can be revolved in a similar manner for any pair of tests we see that the correlation coefficient for any pair of tests is the cosine

of the central angle between them. This is true for any number of factors that may be postulated and hence for any space order.

Our problem can now be restated as follows. We have a table of intercorrelations which are cosines of central angles between pairs of points. Every test is represented by a point on the surface of a sphere. Since all of the correlation coefficients are subject to experimental errors, we cannot trust the accuracy of any one of them. Our problem is to find the best fitting allocations of the points on the surface of the sphere. The least square methods are here altogether too unwieldy. They are out of the question, undoubtedly. We shall therefore apply an adaptation of the principles of curve fitting by the method of averages. Just as in the method of averages we deal with summations of partial sets of observations, so we shall here deal with summations for partial sets of tests or points. In fact, we shall define the $x$-axis so that it passes through the origin and through the center of gravity of a partial set of points. The $y$-axis will be so determined that the $xy$-plane contains the $x$-axis and the center of gravity of a second set of points. This procedure can be continued for any number of dimensions or factors. The number of dimensions of the sphere is the same as the number of general factors that are postulated. This procedure will now be described in more detail.

In locating the coördinate axes there are several criteria and several quantitative tests that may be applied. These criteria will be considered in a separate paper, and we shall here describe only one such criterion by which the axes may be located. Assume that you have before you the sphere with a point designated on its surface for each test or variable in the correlation table. Our first problem is to locate the $x$-axis through this sphere. The projection of a point on this axis will be the loading of the first factor in that test. Suppose that we arbitrarily pass the $x$-axis through test $a$. Then this test will have loading of unity for the first factor and zero for all of the other factors. The projection of test $b$ on this axis will then be the loading of the first factor in

test *b*. In dealing with actual data we shall always have a residual composed of minor general factors that are ignored and specific factors peculiar to each test or variable. For this reason we want to account for as much as possible of the correlations in terms of the smallest possible number of general independent factors. We therefore want to pass the $x$-axis through this sphere so as to maximize the projections of all the points on it.

We first find that one test which has the highest average correlation with all the other tests, disregarding sign. We disregard sign because it does not matter here whether the projection is positive or negative. The sign is arbitrary and either projection serves to determine the correlation. The sum of each column in the correlation table, disregarding sign, is recorded. Let the test with the highest sum be designated test *a* and let the sum be designated $\Sigma|r_{ak}|$. This test is evidently more nearly like the rest of the tests in the series than any other single test. The average distance from it to all the other tests is, in general, smaller than the corresponding average separation of each other test from all the rest. We have found now the most representative single test in the battery. Of course we could now simply pass the $x$-axis through this test and thereby make sure that we have nearly maximized the projections of all the tests on this axis. But we do not want to define any of the axes by any single test or variable. We therefore make a list of all the tests that correlate positively with test *a*. These tests will all lie in a hemisphere with test *a* at its pole. This is clear because all positive correlations correspond to angles less than 90 degrees, and all correlations that are negative correspond to angles greater than 90 degrees. We can now be fairly sure, but not absolutely so, that this partial group of tests have something in common that is a conspicuous and important general factor in the whole set of tests. We shall define the $x$-axis so that it passes through the center of the sphere, of course, and also through the center of gravity of all the tests in the hemisphere with *a* at its pole.

Let us designate all the tests which correlate positively

with test $a$ by the general notation $s$. The notation $s$ refers then to each of the tests in this partial group, taken in succession. The correlation of test $a$ with each of the tests in this sub-group $s$ may then be written as

$$r_{as} = a_1 s_1 + a_2 s_2 + a_3 s_3. \tag{6}$$

Summing, we have

$$\Sigma r_{as} = a_1 \Sigma s_1 + a_2 \Sigma s_2 + a_3 \Sigma s_3, \tag{7}$$

in which $\Sigma s_1$ is the sum of the projections of the points in the set $s$ on the first axis, $\Sigma s_2$ is the sum of the projections of the points in this set on the second or $y$-axis, $\Sigma s_3$ is the sum of the projections on the third or $z$-axis.

The center of gravity of the points in the set $s$ is a point inside the sphere and since we have taken a set of tests in a single hemisphere we are certain that this center of gravity will not lie at or near the center of the sphere. If that should happen our subsequent determinations would be very inaccurate. The coördinates of the center of gravity of the set of points $s$ will evidently be

$$\frac{\Sigma s_1}{N_s}, \quad \frac{\Sigma s_2}{N_s}, \quad \frac{\Sigma s_3}{N_s}$$

with reference to the three axes that are still to be chosen. Now let us pass the first axis, the $x$-axis, through the center of gravity of the set of points $s$. But if this center of gravity lies on the $x$-axis, then it is clear that the $y$- and $z$-coördinates must be zero so that

$$\Sigma s_2 = \Sigma s_3 = 0.$$

The correlation between the test $a$ and any one of the other tests in the set $s$ can then be simplified because the last two terms in equation (7) vanish. The expression for this correlation becomes then

$$\Sigma r_{as} = a_1 \Sigma s_1 \tag{8}$$

or, rewriting this explicitly for the loading $a_1$, we have

$$a_1 = \frac{\Sigma r_{as}}{\Sigma s_1} \tag{9}$$

By analogy with equation (5) we can write the following corresponding equation for the set $s$,

$$(\Sigma s_1)^2 + (\Sigma s_2)^2 + (\Sigma s_3)^2 = \Sigma r_{ss} \tag{10}$$

in which $\Sigma r_{ss}$ is the sum of all the correlation coefficients in the full table of the $s$ tests, including self correlations taken as unity, and recording each coefficient twice in the table since it is symmetrical. But since $\Sigma s_2$ and $\Sigma s_3$ are zero by the location of our $x$-axis, we have the simpler relation

$$(\Sigma s_1)^2 = \Sigma r_{ss} \tag{11}$$

or

$$\Sigma s_1 = \sqrt{\Sigma r_{ss}}$$

and from this we know the loading of the first general factor in each test as follows:

$$\begin{aligned} a_1 &= \frac{\Sigma r_{as}}{\sqrt{\Sigma r_{ss}}}, \\ b_1 &= \frac{\Sigma r_{bs}}{\sqrt{\Sigma r_{ss}}}, \\ &\cdot \quad \cdot \quad \cdot \\ w_1 &= \frac{\Sigma r_{ws}}{\sqrt{\Sigma r_{ss}}}. \end{aligned} \tag{12}$$

It should be noted that these relations are valid even though the test $b$, for example, is not a member of the particular set of points $s$. Hence the above simple relation enables us to determine the loading of the first general factor in each of the $w$ tests.

We now want to determine the loading of the second general factor in each of the tests. Imagine again the sphere with the $w$ points allocated to its surface. We have now passed the $x$-axis through this sphere. The $y$-axis must of course be at right angles to the $x$-axis because we assume that the general factors are uncorrelated. The angle between the first and second axes must therefore be 90 degrees. If we revolve the sphere about the $x$-axis with test $a$ somewhere in the vicinity of the pole, it is clear that the $y$-axis which shall represent the second general factor will pierce the surface of

the sphere somewhere in the equator. We shall now select a pivot test for the second factor which shall serve the same function in locating the second factor that test *a* served in locating the first factor. This second test will clearly lie in an equatorial band and it will have a low correlation with test *a*. We therefore tabulate separately all the tests that have a low correlation with test *a*. In doing so we must decide how wide an equatorial band is to be included. The band should be wide enough to include all tests that are essentially different from test *a*, but the band should not be so wide as to include those tests which are heavily loaded with the first factor which is already represented by the *x*-axis. The width of the equatorial band selected will depend also in part on the number of tests in the whole series of intercorrelations. Let us decide to use an equatorial band not wider than 30 degrees on either side of the equator. An angle of this size has a cosine of .50 so we make a list of all tests that correlate less than .50, either positive or negative, with test *a*.[2]

In order to find the most representative test in the equatorial band, we inspect again the summations $\Sigma|r_{ak}|$, $\Sigma|r_{bk}|$, $\Sigma|r_{ck}|$, $\cdots$ $\Sigma|r_{wk}|$, disregarding sign. We find the test *b* in this list which has the highest sum of its correlations with the other tests. In the equatorial band this test is the most representative of all the tests in the whole series *w*. By selecting this one as our second pivot test we will, in general, but not necessarily, maximize the projections of all the tests on the second axis. Let this second pivot test, so selected, be test *b*.

We now tabulate, as before, all those tests in the whole series *w* which correlate positively with test *b*. Let this second sub-group of tests be designated by the notation *t*. We have now a set of tests *t* which lie in a hemisphere with test *b* as its pole. Consider the center of gravity of this set of points *t*. It will be a point inside the sphere but it will

[2] In case all the intercorrelations in the initial table are positive it is clear that all the points lie in one quadrant or octant of the surface and hence the selection of the subgroups *s*, *t*, *u*, must be adapted to the distribution of the initial coefficients as to positive and negative values.

certainly not coincide with the center of the sphere. That is to be avoided in the interest of accuracy of subsequent determinations. Now we shall locate the $y$-axis so that the $xy$-plane contains the $x$-axis and also the center of gravity of the set of points $t$. By so doing we make the $z$-coördinate of the center of gravity of the set $t$ vanish. By this procedure in locating the successive axes we reduce the number of unknowns so that the loadings may be readily determined.

The correlation between test $a$ and any one of the tests in the set $t$ may be written

$$r_{at} = a_1 t_1 + a_2 t_2 + a_3 t_3. \tag{13}$$

Summing for all the tests in the set $t$, as before, we have

$$\Sigma r_{at} = a_1 \Sigma t_1 + a_2 \Sigma t_2 + a_3 \Sigma t_3. \tag{14}$$

Here the values of $\Sigma t_1$, $\Sigma t_2$, $\Sigma t_3$, are the sums of the projections of the points in the set $t$ on each of the three coördinate axes.

The coördinates for the center of gravity of the set of points $t$ are evidently

$$\frac{\Sigma t_1}{N_t}, \quad \frac{\Sigma t_2}{N_t}, \quad \frac{\Sigma t_3}{N_t},$$

but the $z$-coördinate vanishes by so locating the $y$-axis that the $xy$-plane contains the center of gravity for the set of points $t$. Therefore

$$\Sigma t_3 = 0$$

and hence

$$\Sigma r_{at} = a_1 \Sigma t_1 + a_2 \Sigma t_2 \tag{15}$$

Since all the $x$-coördinates are now known, so is also $\Sigma t_1$ for the set $t$. Therefore we can solve for $a_2$ in the above equation in the form

$$a_2 = \frac{\Sigma r_{at} - a_1 \Sigma t_1}{\Sigma t_2}. \tag{16}$$

The summation $\Sigma t_2$ is known from the relation

$$\Sigma r_{tt} = \tfrac{n}{n}(\Sigma t_1)^2 + (\Sigma t_2)^2, \tag{17}$$

which can be written by analogy with equation (5). This equation enables us to determine the loadings of the second factor in all of the tests.

The loadings of the second factor in the remaining tests are determined by equations analogous to (16), namely

$$b_2 = \frac{\Sigma r_{bt} - b_1 \Sigma t_1}{\Sigma t_2}, \qquad (16)$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot$$

$$w_2 = \frac{\Sigma r_{wt} - w_1 \Sigma t_1}{\Sigma t_2}.$$

When the loadings of the first two factors in each of the tests have been determined it would be possible to determine the loading of the third factor from the relations

$$a_1^2 + a_2^2 + a_3^2 = 1,$$
$$b_1^2 + b_2^2 + b_3^2 = 1,$$
$$c_1^2 + c_2^2 + c_3^2 = 1,$$

and so on provided that we were certain that three factors were sufficient to determine the correlation coefficients. However, we must assume that in any ordinary situation there are residuals composed of additional general factors of minor importance, perhaps, and specific factors peculiar to each of the tests. On this account we shall not make use of the above equation for determining the loadings of the third factor. These loadings will be determined in a manner similar to that used for the previous loadings.

We make a list of all the tests that correlate between + .50 and − .50 with *both* tests *a* and *b*. In this manner we shall be sure that we have a list of tests represented by points which are in the general vicinity of a right angle from both *a* and *b*. We now leave the restricted three dimensional sphere and proceed by analogy into higher dimensions. We select that one test in this list which has the highest sum $\Sigma r_{kk}$, disregarding sign. Let this test be test *c*. This will be the pivot test for the sub-group of tests which are to be used for determining the third general factor.

After having found the pivot test for the third general factor we list all of the tests that have positive correlations with test *c*. Let this sub-group be designated *u*. These will all lie in a hemisphere with test *c* at its pole. We shall

so locate the third axis or factor that the center of gravity of the set of points $u$ has finite values for the first three coördinates, and so that the higher numbered coördinates of this center of gravity will vanish. This is merely carrying out the same procedure as before. The correlation between test $a$ and any one of the tests in set $u$ will then be

$$r_{au} = a_1u_1 + a_2u_2 + a_3u_3. \tag{17a}$$

Summing we have

$$\Sigma r_{au} = a_1\Sigma u_1 + a_2\Sigma u_2 + a_3\Sigma u_3. \tag{18}$$

This equation can be solved for $a_3$ since all the other values are known. The values of $\Sigma u_1$ and $\Sigma u_2$ are known because we have the loadings of the first and second factors in each of the tests and we have listed the tests that belong in the set $u$. The summation $\Sigma r_{au}$ is known directly from the given correlation coefficients. The value of $\Sigma u_3$ can be obtained from the equation

$$\Sigma r_{uu} = (\Sigma u_1)^2 + (\Sigma u_2)^2 + (\Sigma u_3)^2, \tag{19}$$

which is written by analogy with equation (5). Hence

$$a_3 = \frac{\Sigma r_{au} - a_1\Sigma u_1 - a_2\Sigma u_2}{\Sigma u_3}, \tag{20}$$

$$b_3 = \frac{\Sigma r_{bu} - b_1\Sigma u_1 - b_2\Sigma u_3}{\Sigma u_3},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$w_3 = \frac{\Sigma r_{wu} - w_1\Sigma u_1 - w_2\Sigma u_3}{\Sigma u_3}.$$

In the same manner we may extend the procedure to any number of factors that may be necessary to account for a given table of intercorrelations. However, it is not advisable to carry this procedure so far as to determine the correlations within the errors of measurement because in that case the last factors are likely to be merely the loadings that are necessary to adjust for chance errors. Our purpose is to discover only those principal factors that are truly operative in producing the correlation coefficients. Hence this procedure should be carried out far enough to lock the coefficients

with discrepancies somewhat greater than the chance errors in the given coefficients.

Finally, when the loadings have been determined for each test, one should calculate all of the coefficients by equations (2). These can then be compared with the given experimental coefficients. The discrepancies $d = r_e - r_c$ in which $r_c$ = calculated $r$, and $r_e$ = experimental or observed $r$, should then be calculated. A frequency distribution of these discrepancies will indicate fairly well how closely the factors account for the given coefficients. In no case should one expect the standard deviation of this distribution to be the same as the average standard error of the given coefficients. The standard deviation of the discrepancies will be the larger because it may be generally assumed that the limited number of general factors that are postulated do not include the minor group factors and specific factors that together will make the residuals greater than those expected by chance errors alone.

In order to facilitate the application of the procedure we shall list here the successive steps together with the fundamental equations that are necessary in solving for the successive values.

*The following outline has been extended to include four factors in order to illustrate the method.*

(1) Prepare a full table of intercorrelations for the $w$ tests. Assume that all self correlations are unity and record them so. The full table will be symmetrical in that every coefficient will occur in two cells. Thus the coefficient $r_{ab}$ will occur in column $a$ opposite $b$ and also in column $b$ opposite $a$. Record the sums $\Sigma|r_{ak}|$, $\Sigma|r_{bk}|$, $\Sigma|r_{ck}|$, etc., for all columns, disregarding sign. The subscript $k$ refers to each of the $w$ tests in the whole series taken in succession.

(2) Find that test $a$ which has the highest sum for its column, namely $\Sigma|r_{ak}|$.

(3) Make a list of all the tests that correlate positively with test $a$. This is done by listing all tests with positive correlations in column $a$. Let this sub-group of tests

be designated $s$. The sum $\Sigma r_{ss}$ is the *algebraic* sum of all the coefficients in a full table of the tests in group $s$. Then

(4) $$(\Sigma s_1) = \sqrt{\Sigma r_{ss}}$$

(5) $$a_1 = \frac{\Sigma r_{as}}{\sqrt{\Sigma r_{ss}}}$$

$$b_1 = \frac{\Sigma r_{bs}}{\sqrt{\Sigma r_{ss}}}$$

$$\cdot \quad \cdot \quad \cdot$$

$$w_1 = \frac{\Sigma r_{ws}}{\sqrt{\Sigma r_{ss}}}$$

The loading of the first factor in every test is now determined.

(5*a*) Determine the algebraic sum $\Sigma s_1$ from the list of first factor loadings and see that it agrees with the value found in step 4.

(6) Make a list of all tests that correlate with test $a$ within the range $\pm .50$. These tests are in general different from test $a$. Select that test $b$ in this list which has the highest sum in its column, namely $\Sigma|r_{bk}|$, disregarding sign.

(7) Now make a list of all the tests that correlate positively with test $b$ and let this sub-group be designated $t$.

(8) Determine $(\Sigma t_1)$. This can be done since we know the first factor loading in each test and we have a list of the tests in group $t$. Also determine $\Sigma r_{tt}$ for the sub-group $t$.

(9) Determine

$$\Sigma t_2 = \sqrt{\Sigma r_{tt} - (\Sigma t_1)^2}.$$

(10) Determine

$$a_2 = \frac{\Sigma r_{at} - a_1 \Sigma t_1}{\Sigma t_2},$$

$$b_2 = \frac{\Sigma r_{bt} - b_1 \Sigma t_1}{\Sigma t_2},$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot$$

$$w_2 = \frac{\Sigma r_{wt} - w_1 \Sigma t_1}{\Sigma t_2}.$$

(10*a*) Determine the sum $\Sigma t_2$ from the list of second factor loadings and make sure that it agrees with value found in step 9.

(11) Make a list of all tests that correlate low, say within the range $\pm .50$, with *both* tests *a* and *b*. If there are no such tests, then you need no additional general factors and this procedure need not be carried any further. Select that one test *c* in this restricted list that has the highest sum $\Sigma|r_{ck}|$, disregarding sign.

(12) Make a list of all tests that correlate positively with test *c* and let this sub-group be designated *u*.

(13) Determine $\Sigma u_1$ and $\Sigma u_2$. This can be done since we have the first and second factor loadings for each test and we have a list of all tests in the sub-group *u*. Also determine $\Sigma r_{uu}$ which is the algebraic sum of all coefficients in the full table for group *u*.

(14) Determine

$$\Sigma u_3 = \sqrt{\Sigma r_{uu} - (\Sigma u_1)^2 - (\Sigma u_2)^2}.$$

(15) Determine

$$a_3 = \frac{\Sigma r_{au} - a_1\Sigma u_1 - a_2\Sigma u_2}{\Sigma u_3},$$

$$b_3 = \frac{\Sigma r_{bu} - b_1\Sigma u_1 - b_2\Sigma u_2}{\Sigma u_3},$$

. . . . .

$$w_3 = \frac{\Sigma r_{wu} - w_1\Sigma u_1 - w_2\Sigma u_2}{\Sigma u_3}.$$

(15*a*) Determine the sum $\Sigma u_3$ from the list of third factor loadings and make sure that it agrees with the value found in step 14.

(16) Make a list of all tests that correlate low, say within the range $\pm .50$, with the tests *a*, *b*, and *c*. If there are no such tests, then you need no additional factors. If there are several such tests, continue as follows. Select that one test *d* in this restricted list that has the highest sum $\Sigma|r_{dk}|$, disregarding sign.

(17) Make a list of all tests that correlate positively with test *d* and let this sub-group be designated *v*.

(18) Determine $\Sigma v_1$, $\Sigma v_2$, and $\Sigma v_3$. This can be done since we have the first, second, and third factor loadings in each of the $w$ tests and we have a list of the tests in sub-group $v$. Also determine the algebraic sum $\Sigma r_{vv}$.

(19) Determine

$$\Sigma v_4 = \sqrt{\Sigma r_{vv} - (\Sigma v_1)^2 - (\Sigma v_2)^2 - (\Sigma v_3)^2}.$$

(20) Determine

$$a_4 = \frac{\Sigma r_{av} - a_1\Sigma v_1 - a_2\Sigma v_2 - a_3\Sigma v_3}{\Sigma v_4},$$

$$b_4 = \frac{\Sigma r_{bv} - b_1\Sigma v_1 - b_2\Sigma v_2 - b_3\Sigma v_3}{\Sigma v_4}$$

. . . . . .

$$w_4 = \frac{\Sigma r_{wv} - w_1\Sigma v_1 - w_2\Sigma v_2 - w_3\Sigma v_3}{\Sigma v_4}.$$

It is evident that this procedure can be extended in cycles of five steps for each new factor. Thus a new factor is started in steps 1, 6, 11, 16, 21, and new factors would appear in steps 26, 31, 36, and so on, if the process were continued. The process should be continued until no new tests appear in the lists prepared in each of the steps just enumerated.

In this manner one can make some rational estimate of the number of factors that will be needed to account for any given table of coefficients. Note that this estimate can be made by preparing the lists of tests called for in steps 6, 11, 16, 21, 26, 31, and so on until the list vanishes, before any calculations are started. The number of factors can thus be estimated merely by inspection of the coefficients and without any calculating whatever. This enables one to lay out in advance the data sheets for a postulated number of general and independent factors.

As stated at the outset there are several other criteria that may be applied in selecting the general factors or coördinate axes which are of importance in special cases where unusual clustering of the tests or variables may be suspected. Some special methods of locating the coördinate axes will be considered in a separate paper.

We have described a method of multiple factor analysis

by which it is possible to ascertain how many general, independent, and uncorrelated factors it is necessary to postulate in order to account for a whole table of intercorrelations. The method is free from any limitations about group factors. Objective methods have been described for locating the general factors and it is probable that, except in unusual distributions of the tests, these methods will prove adequate.

After the factor loadings have been determined for all of the tests or variables, it is of considerable interest to describe or name the general factors that have been isolated. This can be done best by noting which tests or variables have a relatively high positive loading with the first factor and which have a low or negative loading with this factor. By this inspection it is possible to name the first factor although the statistical procedures do not of course concern these matters of describing or naming the factors. In a similar manner one might inspect the second factor loadings to ascertain which tests have positive loadings and which have negative loadings with this factor. A name might then be found for this factor and so on.

Finally, it may be of interest to assign the standard score in each of the factors to each of the individual subjects. If each factor represents an ability of some kind, then it would be of interest to be able to assign to each person a standard score in each of the factors or abilities that have been isolated. In the following list of equations the only unknown factors are $x_1$, $x_2$, $x_3$, and $x_4$. These standard scores in the four abilities for any one person may be determined by solving the following observation equations for the four unknown standard scores by the method of least squares or by the method of averages. This procedure is laborious but it is at least possible to solve the problem of individual standard scores in the several independent abilities.

$$
\begin{aligned}
S_a &= a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4, \\
S_b &= b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4, \\
S_c &= c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4, \\
&\cdots \\
S_w &= w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4.
\end{aligned}
$$

It is probable that these methods of multiple factor analysis will be useful in discovering how many factors underlie a given table of correlation coefficients and in discovering their general nature. Several applications of the method to different types of correlation problems are under way and these will be reported in subsequent papers.

[MS. received March 9, 1931]