STATISTICAL METHODS FOR PSYGHOLOGY AND EDUCATION

BY DR RAJIV SAKSENA

DEPARTMENT OF STATISTICS

UNIVERSITY OF LUCKNOW

INTRODUCTION

Psychometry is the branch of psychology which deals with the measurement of psychological traits or mental abilities like intelligence, aptitude, interest, opinion, attitude or, simply scholastic achievement. Education statistics may be considered to be a part of psychiatry where our main purpose is to rank a group of individuals according to their scholastic achievement. Although this task of ranking does not seem to present immediate problems, a close examination will reveal a number of pitfalls and weaknesses of the prevalent system.

Unlike physical or biological characteristics, psychological characteristics are rather abstract and hence can be measured only with some degree of unreliability. For the purpose of measurement, one has to develop a certain scale, which bears a strong analogy with a foot-rule used for measuring or comparing lengths. As on a foot–rule equal distances on a psychological scale stand for empirically equal differences n the psychological trait being measured. But the zero–point of the psychological scale, unlike that of the foot-rule, is arbitrary. However, distances from the arbitrary zero are additive. In other words a psychological scale is an *interval scale* and not a *ratio scale,* since there is no absolute zero-point on it.

SOME SCALING PROCEDURES

Most of the scaling procedures used for psychological or educational data are based on the assumption that the trait under consideration is normally distributed. The zero-point and the units of the scale are chosen arbitrarily, but the scale unit should be equal and remain stable throughout the scale. We shall discuss in this section some of the common scaling procedures used in psychology and education.

SCALING INDIVIDUAL TEST-ITEMS OF DIFFICULTY

Here we have a number of items in a test administered to a large group of individuals. The proportion of individuals successful in each item is known. We assume in the construction of the difficulty scale that the ability (x) which the group of items is measuring is normally distributed with some mean μ and some s.d σ. We can arbitrarily take the origin at μ write μ=0.

Let p_i be the proportion of individuals passing the ith item. We determine the point x-axis for which the area to the right of the ordinate is p_i . Let the point be $k_i\sigma$. Thus $k_i\sigma$ is the amount of ability required for passing the item and hence may be taken as a measure of difficulty (d_i) for the ith item. Thus an equal difference in d will mean an equal difference in ability required for passing the items.

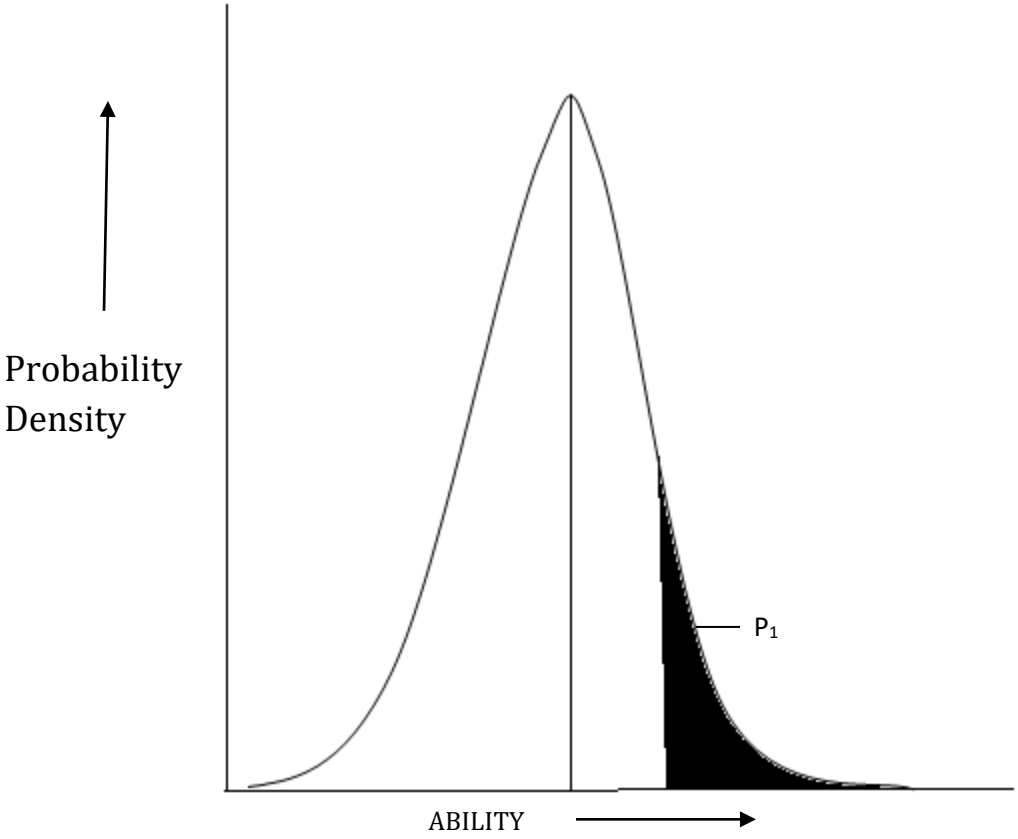


Fig. Determining the difficulty-value of an item from the proportion of individuals passing the item

Example: Suppose there are four item A,B,C and D, passed, respectively, by 90%, 80%, 70% and 60% of the individuals. Compare the difference in difficulty between A and B with the difference in difficulty between C and D.

To find the difficulty value d_A of the item A we find the point, on the normal distribution with mean 0, and s.d. σ, the area to the right of which is 0.90. From the table of the area under the normal probability curve (Table I, Appendix B), we have

$$d_A = -1.28\sigma$$

Similarly $$d_B = -0.84\sigma$$

$$d_C = -0.52\sigma$$

And $$d_D = -0.25\sigma$$

Hence $$d_B - d_A = 0.44\sigma, \quad \text{whereas} \quad d_D - d_C = 0.27\sigma$$

Thus $$\frac{d_B - d_A}{d_D - d_C} = \frac{0.44\sigma}{0.27\sigma} = 1.63.$$

The difficulty of B relative to A is 1.63 times greater than the difficulty of D relative to C.

SCALING OF TEST-SCORES IN SEVERAL TESTS

The main defect of the prevalent system of ranking in scholastic test consists in the adding of the raw scores of an individual on several tests to get his composite or total scores and ranking all individuals on the basis of the total scores. This is not a valid procedure since the same raw score x on different tests may involve different degrees of ability and hence may not be equivalent in different tests. Hence the raw scores have to be scaled under some assumption regarding the distribution of the trait which the test is measuring.

Percentile scaling

Here we assume that the distribution of the trait under consideration is rectangular, under which we shall have percentile differences equal throughout the scale. To determine the scale value corresponding to a score x we have to find the percentile position of an individual with score $x, i.e$ the percentage of individuals in the group having a score equal to or less than x, which can be easily obtained from the score-distribution assuming that 'score' is a continuous variable .Regardless of the form of the original raw scores distribution, the distribution of percentile score will be rectangular. However, the distribution of raw scores is rarely rectangular, so that the basic assumption underlying the percentile scaling may not always be realistic. Thus while using this scaling method one should beware of its limitations.

Z-Scaling or σ scaling

Here we assume that whatever differences that may exist in the forms of raw score distributions may be attributed to chance or to the limitations of the test. In fact, the distributions of the traits under consideration are assumed to differ only in mean and s.d. Hence the score on different tests should be expressed in items of the score in a hypothetical distribution of the same form as the trait–distribution with some arbitrarily chosen mean and s.d. The transformed scores are called *linear derived scores*. In particular, if the mean is arbitrarily taken to be zero and the s.d to be unity, the scores are called *standard scores or σ-scores or z–score*. To avoid negative standard scores, in linear derived scores the mean is generally taken to be 50 and the s.d to be 10. If a particular test has raw score mean & s.d equal to μ & σ, respectively, then the linear derived score corresponding to a score x on that test is given by

$$\frac{x-\mu}{\sigma}=\frac{w-50}{10}$$

or
$$w = 50 + 10X\frac{(x-\mu)}{\sigma} = 50 + 10z, \qquad \ldots(1)$$

where w is the linear derived score with mean 50 and s.d. 10 and z is the standard score. This linear transformation changes only the mean and the s.d., while retaining the form of the original distribution.

T-Scaling

In this case we assume that the trait-distribution is normal. The raw score distribution may deviate from normality, but the deviations from normality are attributed to chance or to limitations of the tests. The mean and s.d. of the normal distribution of the trait may be arbitrarily taken to be 50 and 10, respectively. To get the scaled score corresponding to a raw score x, first we find as in percentile scaling, the percentile position (P) of an individual with score x and then find the point (T) on a normal distribution with mean 50 and s.d. 10 below which the area is P/100. This is given by.

$$\Phi\left(\frac{T-50}{10}\right)=\frac{P}{100} \qquad \ldots(2)$$

Where $\Phi(r)$ is the area under the curve of the normal deviate from -∞ to r.

The scaled score obtained by the process is called T–score in memory of the psychologists Terman and Thorndyke. The scale is due to McCall.

Normalized scores are also expressed as stanine (standard nine) score. The stanine scale takes nine values from 1 to 9, with mean 5 and s.d 2. When a distribution is transformed to a stanine scale, the frequencies are distributed as follows:

TABLE 5.1

STANINE DISTRIBUTION

Stanine score	1	2	3	4	5	6	7	8	9
Percentage on each score (rounded)	4	7	12	17	20	17	12	7	4

A transformation is nonlinear if it changes the form of the distribution. Normalized score and percentile score are merely special case of nonlinear transformation of the raw score. For nonlinear transformation any form of distribution may be chosen.

Method of equivalent scores

Here we do not make any assumption about the distribution of the trait under consideration. The appropriate trait distribution is obtained by graduating the raw score distribution by an appropriate Pearsonian curve.

Let x and y be the scores on two tests, having probability–density function $f(x)$ and $h(y)$, respectively, obtained by some process of graduation. Now, two score on the two tests x_i and y_i are to be considered equivalent, in the score that they bring into play equal amounts of the trait, if and only if

$$\int_{-\infty}^{x_i} f(x)dx = \int_{-\infty}^{y_i} h(y)dy. \qquad \ldots(3)$$

For Practical convenience an equivalence curve may be obtained by computing a number of pairs of equivalent score, $(x_i y_i)$ and fitting to the corresponding set of points an appropriate curve, say $y = g(x)$.

Equivalent score can also be obtained from the score distribution for x and y without going into the process of graduation .First two ogives are drawn on the some graph paper. Two scores x_i and y_i with the same relative cumulative frequency are then regarded are equivalent.

For the purpose of comparison or combination, the raw score on different tests may be converted into equivalent scores on a standard test. In This method the form of the distribution of equivalent (transformed) scores is the same as that of the standard test. If however, the standard test score has a normal distribution, the method reduces to normalized scaling.

Example: The raw score distribution for Vernacular and English for a group of 500 students are given below. One of two students got 80 in Vernacular and 40 in English, while the other got 60 in both. Compare their performances by (i) percentile scaling, (ii) linear derived scores, (iii) T-scaling and (iv) equivalent score (ogive method).

First we have to remember that a score of 80 is to be considered as an interval from 79.5 to 80.5, and similarly for the other scores. To obtain the percentile positions, we obtain the cumulative frequencies (less-then type) for both Vernacular and English. They are shown in Table. Hence the percentile positions, corresponding to 80.5 and 60.5 in vernacular are given by

$$P_{80.5}(Vern) = \frac{497+0.6}{500} \text{ X } 100 = 99.52$$

And $$P_{60.5}(Vern) = \frac{436+7.2}{500} \text{ X } 100 = 88.64$$

Similarly, for English $$P_{40.5}(Eng.) = \frac{270+15.6}{500} \text{ X } 100 = 57.12$$

and $$P_{60.5}(Eng.) = \frac{476+3.6}{500} \text{ X } 100 = 95.92.$$

TABLE: DISTRIBUTIONS OF SCOBES IN VERNACULAR AND ENGLISH OF A GROUP OF 500 STUDENTS

Score	*Frequency*	
	Vernacular	*English*
0-4		*3*
5-9		*6*
10-14		*12*
15-19	*6*	*23*
20-24	*7*	*35*
25-29	*18*	*45*
30-34	*34*	*74*
35-39	*56*	*72*
40-44	*84*	*78*
45-49	*74*	*53*
50-54	*104*	*46*
55-59	*53*	*29*
60-64	*36*	*18*
65-69	*16*	*5*
70-74	*9*	*1*
75-79	*0*	
80-84	*3*	

TABLE: CUMULATIVE FREQUENCY DISTBIBUTIONS OF SCORES IN VERNACULAR AND ENGLISH

Score	*Cumulative Frequency*	
	Vernacular	*English*
0-4		*3*
5-9		*9*
10-14		*21*
15-19	*6*	*44*
20-24	*13*	*79*
25-29	*31*	*124*
30-34	*65*	*198*
35-39	*121*	*270*

40-44	*205*	*348*
45-49	*279*	*401*
50-54	*383*	*447*
55-59	*436*	*476*
60-64	*472*	*494*
65-69	*488*	*499*
70-74	*497*	*500*
75-79	*497*	
80-84	*500*	

Hence the total scaled score for student 1, getting 80 in Vernacular and 40 in English, is by percentile scaling,

99.52+57.12= 156.64

And that of student 2, getting 60 in both Vernacular and English, is

88.64 +95.92=156.64

Thus we see that the relative performances of the two students are quite different although their total raw scores are equal.

For linear derived scores with mean 50 and s.d. 10, we require the means and s. d.s of scores in the two subjects. Denoting by x the score in Vernacular and by y the score in English, we have.

$\bar{x}$ =47.07 $s_x = 11.32$

$\bar{y}$ =37.87 And $s_y = 13.10$

Hence the w scores are given by

$$w_{80}(Vern) = 50 + \frac{80 - 47.09}{11.32} X10 = 79.07,$$

$$w_{80}(Vern) = 50\frac{60 - 47.09}{11.32} X10 = 61.40,$$

$$w_{80}(Eng) = 50 + \frac{40 - 7.87}{13.10} X10 = 51.63$$

and $$w_{80}(Eng) = 50 + \frac{60-37.87}{13.10} X10 = 66.89.$$

As such, the total w −score of student 1 is

$$79.07+51.63=130.70,$$

and the of student 2 is

$$61.40+66.89 =130.70,$$

Linear derived scores however show that student I is slightly superior to student 2.

Now, for T-scaling percentile positions have to be converted into T-score. We have

$$T_{80}(Vern) = 50 + r_{.9952} X10 = 75.90,$$

$$T_{60}(Vern) = 50 + r_{.8864} X10 = 62.08$$

$$T_{40}(Vern) = 50 + r_{.5712} X10 = 51.79$$

and $$T_{60}(Vern) = 50 + r_{.9592} X10 = 67.41$$

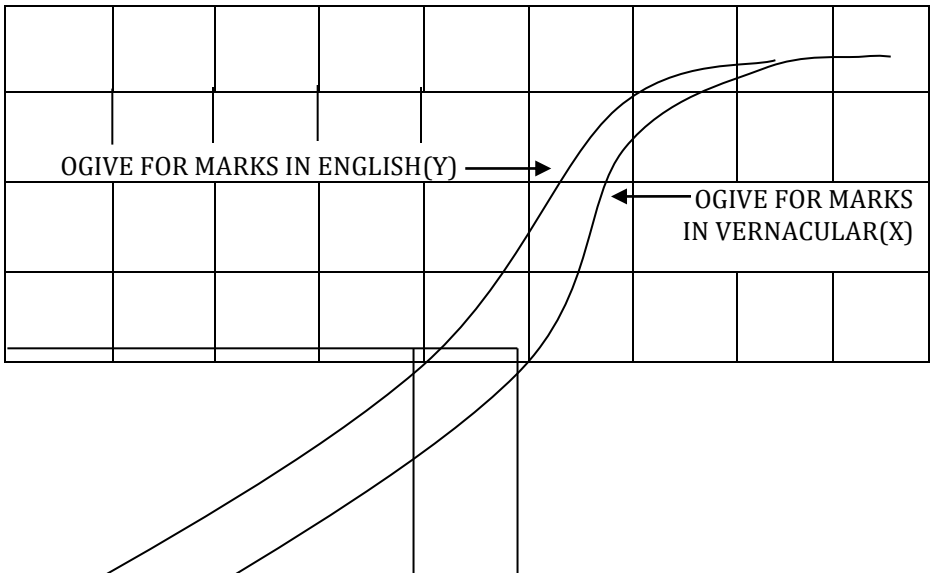
Hence the total T-score of student 1 is

$$75.90+51.79=127.69$$

and the total T-score of student 2 is

$$62.08+67.41=129.49$$

Thus T-scaling shown that student 2 is slightly superior to student 1



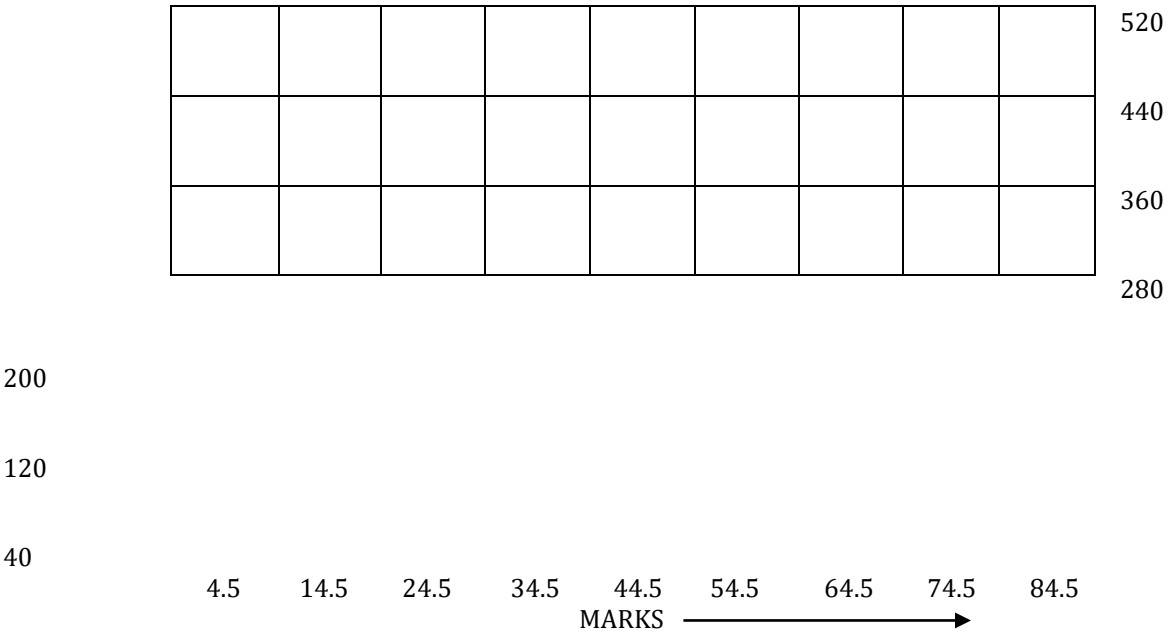


Fig. Determination of equivalent scores in English
and Vernacular from the ogives

In the equivalent score method, let us take Vernacular as the standard. From figure we find that a score of a 40 in English is equivalent to a score of 49.8 in Vernacular and a score of 60 in English is equivalent to a score of 66.9 in Vernacular.

Hence the total score of student 1 in terms of Vernacular score is

80+49.8=129.8

And that student 2 is

60+66.9=126.9

This method again shows that student 1 is slightly superior to student 2.

SCALING OF RATING OR RANKING TO TERM OF NORMAL CURVE

In many psychological problems, individuals are rated or ranked by judges for their possession of some characteristics not readily measurable in terms of performance. Honestly responsibility tactfulness etc, are examples of such traits. Suppose that there are two judges rating a group of individuals

and that the frequency distributions of ratings for the two judges are known. The problem is to assign 'weights' or numerical scores to the ratings, so that the ratings of the judges may be compared or combined.

Let us assume that the distribution of the trait ($say\ x$) is normal with mean 0 and s.d. 1. Now suppose that the individuals with trait values from x_1 to x_2 are given a particular rating. The scale value for the rating is taken to be the mean trait value of all these individuals and so is given by the formula:

$$\text{Scale value} = \frac{\int_{x_1}^{x_2} x \frac{1}{\sqrt{2\pi}} = \exp[-x^2/2]dx}{\int_{x_1}^{x_2} x \frac{1}{\sqrt{2\pi}} = \exp[-x^2/2]dx}$$

$$= \frac{\left[-\frac{1}{\sqrt{2\pi}}\exp[-\frac{x^2}{2}\right]_{x_1}^{x_2}}{\Phi(x_2) - \Phi(x_1)} = \frac{\Phi(x_1) - \Phi(x_2)}{\Phi(x_2) - \Phi(x_1)} \qquad \ldots(4)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp[-x^2/2]$ and $\Phi(x) = \int_{-\infty}^{x} exp[-\mu^2/2]\, d\mu$.

From the observed distribution of the ratings, it is easy to find $\Phi(x_1)$ and $\Phi(x_2)$, and hence $\phi(x_1)$ and $\phi(x_2)$,

The method is due to Likert and the scale is known as *Likert's* scale. This is also called the *category–scale method.* If on the other hand the n individuals in the group are ranked by different judges, the scale values corresponding to the ranks can be obtained under the sane assumptions as before, $i.e$ under the assumption of normality of the trait concerned.

Suppose there is no tie. Then *percentile rank (PR)* of an individual with rank R, $i.e$ percentage of individuals who are ranked below him, is given by

$$PR = 100 - \frac{100(R-\frac{1}{2})}{n} = P, say, \qquad \ldots(5)$$

since the rank R of the individual really represents the interval from $R - \frac{1}{2}$ to $R + \frac{1}{2}$. The scale value corresponding to this PR can now be obtained as the value of a normal deviate below which the area is P/100. In the case of tied ranks, the PR values can be obtained from the frequency distribution of ranks.

Example A group of 100 workers was rated by a supervisor on a five-point scale –A, B,C,D and E–with respect to efficiency, A being the highest rating and E the lowest. Obtain the scale value for each rating from the following frequency distribution of the rating: Obtain the Scale value for each rating from the following frequency distribution of the rating:

Rating	A	B	C	D	E
Frequency	5	24	45	23	3

Under the usual assumption of normality for the trait under consideration, we obtain, for the rating the scale values as follows:

Raking	A	B	C	D	E
Area covered by the rating $\phi(x_2)$- $\phi(x_1)$	0.05	0.24	0.45	0.23	0.03
Area covered by the rating $\phi(x_1)$	0.95	0.71	0.26	0.03	0
Lower limit of the trait x_1	1.645	0.553	-0.643	-1.881	$-\infty$
Upper limit of the trait x_2	∞	1.645	0.553	-0.643	-1.881
Ordinate at the upper limit $\phi(x_1)$	0.1031	0.3424	0.3244	0.0680	0
Ordinate at the upper limit $\phi(x_2)$	0	0.1031	0.3424	0.3244	0.068
Scale value $\frac{\phi(x_1)-\phi(x_2)}{\phi(x_2)-\phi(x_1)}$	2.062	0.997	-0.040	-1.115	-2.267

SCALING OF QUALITATIVE ANSWERS TO A QUESTIONNAIRE

The answers to the items in an attitude or personality test or a test of a similar type will be qualitative $e.g.$, 'Yes ' and 'No', or 'Strongly approve', 'Approve', 'Undecided', 'Disapprove' and 'Strongly disapprove'. It is necessary to allot numerical score to the answers so as to obtain the total score of an

individual measuring his attitude or personality. The method of scaling is exactly similar to Likert's rating scale. The questionnaire is first administered to a group of individuals and the frequency distribution of the answers is obtained. From the observed distribution, Likert's scale values are then obtained for different answers to the questionnaire.

SCALING OF JUDGMENTS OF A NUMBER OF PRODUCTS: PRODUCT SCALE

It often happens that the ability or the trait in which we are interested cannot be expressed as a test score. This necessitates the construction of product scales. In such scales Excellence of performance is determined by comparing an individual's product with various standard products, the values of which are already determined by a number of competent and expert judges, hand-writings, compositions, etc., are well-known examples.

We shall discuss the method of paired comparisons due to Thurston, suppose there are k standard products judges, by a group of N judges. All possible pairs of products $k(k-1)/2$ in all are presented to a judge and he is to select one member of each pair in preference to the other. The can be presented in the form of a proportion matrix.

	Product 1	2	...	....	k
1	p_{11}	p_{21}	....	....	P_{k1}
Product 2	p_{12}	p_{22}	....	....	P_{k2}
:	:	:	:	:	:
k	p_{1k}	p_{2k}	....	....	P_{kk}

Here P_{ij} is the proportion of judges preferring the ith product to the jth one and P_{ji}=1-P_{ij}. By convention P_{ij}=1/2.

Now, suppose that the distribution of difference in judgments (T) of the ith and jth products is normal with mean S_i-S_j (the difference of their scale values) and s.d. σ_{i-j}. Thus

$$P_{ij} = \frac{1}{\sigma_{i-j}\sqrt{2\pi}} \int_o^{\infty} exp\left[-\frac{\{T-(S_i - S_j)\}^2}{2\sigma_{i-j}^2}\right] dT$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-(Si-Sj)/\sigma_{i-j}}^{\infty} \exp[-\tau^2/2] d\tau,$$

so that $S_i - S_j = -x_{ij}\sigma_{i-j}$ ...(6)

where x_{ij} is the value of the normal deviate the area to the right of which is P_{ij}. Equation (5.6) is known as Thurnstone's *law of comparative judgment*. Assuming that the distribution of judgments for each product has the same s.d. σ and that judgments for any two products are uncorrelated, $\sigma_{i-j} = \sigma\sqrt{2}$, a constant.

Taking $\sigma_{i-j} = \sigma\sqrt{2}$ as the unit of the scale, we have

$S_i - S_j = -x_{ij}$ ...(6a)

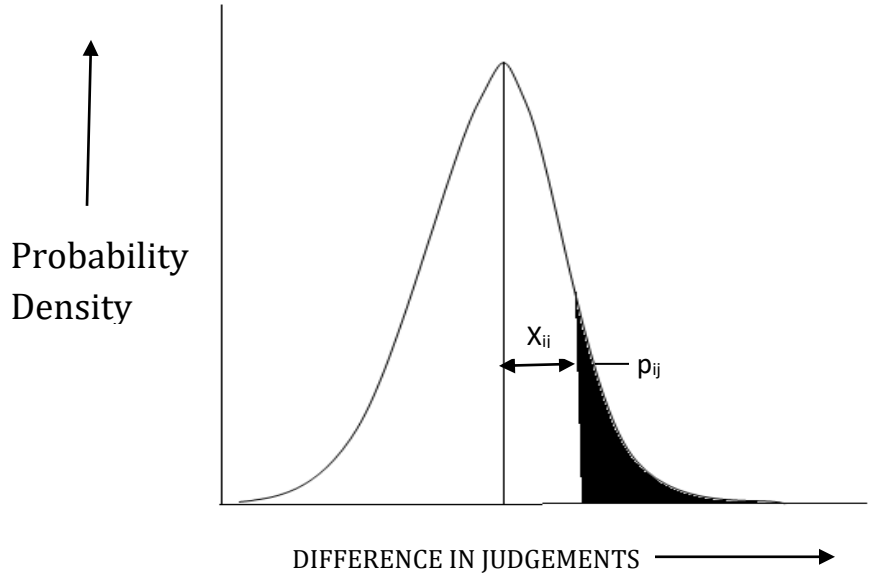


Fig. Determination the difference of scale-value of judgements ($S_i - S_j$) from the proportion p_{ij}

Thus we get the ($S_i - S_j$) matrix:

	1	2	... Product	k

1	s_1- s_1 s_2-s_1 .... s_k-s_1
Product 2	s_1-s_2 s_2-s_2 .... s_k-s_2
:	: : : : :
k	s_1-s_k s_2-s_2 ... . s_k- s_k

The column means give $S_1, S_2, \ldots \ldots S_k$, as deviations from $\bar{S} = \frac{1}{k}\sum_{i=1}^{k} S_i$. If we take the origin at $\bar{S}$, then the column means provide us with the scale – values for the k products. Alternatively, we could take the origin at the minimum scale value and adjust the scale values accordingly.

Example. 200 individuals were asked about their preferences for 4 different types of music. The proportion matrix is given below. Find the scale values.

		Music type			
		1	2	3	4
Music type	1	.500	.770	.878	.892
	2	.230	.500	.743	.845
	3	.122	.257	.500	.797
	4	.108	.155	.209	.500

Under the usual assumption of normality of the distribution of difference in judgments with means $S_i - S_j$and s.d, $\sigma_{i-j,}$and with the constant $\sigma_{i-j,}$ taken as the scale we get the matrix of scale separations $S_i - S_j$ as follows:

		Musictype			
		1	2	3	4
Music type	1	0	.739	1.165	1.237
	2	-.739	0	.653	1.015
	3	-1.17	-.653	0	.831
	4	-1.24	-1.015	-.831	0
Column mean		-785	-.232	.247	.771

With the origin at S the mean scale values, the column means give us the corresponding scale values for the four music types. With origin at S_1, on the other hand, we get the following scale values:

Music type	1	2	3	4
Scale value	0	.553	1.032	1.556

NORMS AND REFERENCE GROUPS

By liner transformation or normalization of test scores, we get the scale values with which we can combine the performances of an individual in different test or can make comparison between individuals. But in many situations, it is not sufficient to have the scale value, but we have to know on the basis of which group of individuals the scaling was done. We have to know the age, sex, education and occupation and other characteristics of the reference group. A scale values with reference to a certain group may not be so good, but it may be very good for another reference group .Thus, when we want to judge the performance of an individual by his test score, we must know what to compare it with,$i.e$ .the norm we want to use. We must know the mean, standard deviation and percentile values for the group with which we compare an individual score. Thus a score may be good when compared to one norm (for a certain reference group).

Many tests are used for several purposes and for several groups of individuals. If the result of a test is to be used for comparison with several groups, it is necessary to have norms for each of the groups separately, unless they are known to be the same. To calculate the norms for several groups, the test has to be administered to a random representative sample from the population of the reference group. The size of the sample should not be too small so as to obtain stable norms .Norm data are however not necessary in practical situations where we want to select a number of individuals out of all applicants on the basis of test score, because the top individuals are to be selected, no what the norms are

TEST THEORY

The measurements on the psychological characteristics considered in previous sections were collected by various types of methods such as tests,

questionnaires or ratings. Whatever may be the method of obtaining measurements, we made the assumption, though not explicitly, that the measurements were meaningful and reproducible. To be more exact we assumed that the measuring instrument used would give us a stable and consistent measure of the trait if we remeasured the trait under identical conditions .Technically, this aspect of the accuracy is known as the ***reliability*** of the measuring instrument. The second requirement is that the measuring instrument measures the trait which it is intended to measure. And technically this is known as the ***validity*** of the measuring instrument.

With physical measurements these present no problems at all. For we know that if we use a non-flexible accurate measuring tape in the correct way, we shall get the exact length of an object, and this can be reproduced if remeasured under similar conditions. So physical measurements are, usually, always reliable and valid. But we are not so sure about psychological measurements. We have to verify in each case that we are getting reliable and valid measurements and then only can we use them with confidence.

Before we actually discuss reliability and validity, we shall consider some simple results in test theory under a very simple model.

LINEAR MODEL OF TEST THEORY

We are interested in getting the true measure of an individual's performance on at test. By applying a measuring instrument what we get is the individual's raw score (obtained score) on the test. We can consider various types of relationship between the true score of the ith individual(t_i) and his raw score (x_i). But the relationship that is usually adopted is the simplest one-a linear relationship. We assume that

$$x_i = t_i + e_i, \quad for\ i = 1,2 \dots \dots \dots \dots, n, \qquad \dots(7)$$

where $e_i = x_i - t_i$ is the error of measurement for the ith individual?

The raw score (x) dose not equal the unknown true score (t). The difference ($x - t$) which may be due to various factors is the error score (e).

In test theory we always consider only random errors(e). Constant or systematic errors are assumed to be absent in test theory. Since we consider only random errors, it is reasonable to make the following assumption for *e's*:

$$\left.\begin{array}{c}\mu_e = 0 \\ \rho_{ie} = 0 \\ \rho_{e_g e_h} = 0\end{array}\right\} \qquad \ldots(8)$$

In words, the mean of error score is zero, the correlation between true score and error score is zero, and the correlation between error score from different testing occasions (or for two parallel tests, g and h, to be defined shortly) is zero. We note that under this model the estimates of μ_e, ρ_{ie} and $\rho_{e_g e_h}$ will approach zero if the number of individuals (n) approaches infinity. In practice, however, the estimates are assumed to satisfy these relations for the given sample.

Since only random errors are considered, for a large number of cases ($n\ large$), the positive and negative errors of all magnitudes (small and large) will cancel each other with the result that the mean will be zero. Similarly, since only random errors are considered, there is no reason to expect any correlation between true scores and error scores for a large number of individuals. Large or small true scores will be expected to occur equally often with large or small error score. This is reasonable for both positive and negative scores. Thus we assume $\rho_{ie} = 0$ A similar argument will show that $\rho_{e_g e_h} = 0$ is also a reasonable assumption.

DEFINITION OF PARALLEL TEST

Two tests are said to be ***parallel*** when it makes no difference which one is used. If g and h are two tests and if for the ith individual $t_{ig} \neq t_{ih}$, then we cannot say that it makes no difference whether we use test g or h. So, in order that g and h may be parallel test, it is reasonable the assume that

$$t_{ig} = t_{ih}, \qquad for\ i = 1,2, \ldots \ldots, n; \qquad \ldots(9)$$

$i.e$, the true score of any individual should be the same on the two tests.

Next, consistent with the definition of error score (8), we assume about the error scores on two parallel tests that

$$\sigma_{e_g} = \sigma_{e_h}; \qquad \ldots(10)$$

$i.e$, the standard deviations of errors on the two tests should be the same. Thus (9) and (10) defined parallel tests in terms of unknown quantities. These can be expressed in terms of the distributions of the raw score, using the relations (7), (8) and (9) as follows:

From (7), since $\mu_e = 0$, we have $\mu_t = \mu_x$ for any test.

From (9), we have $\mu_{t_g} = \mu_{t_h}$, $\sigma_{t_g} = \sigma_{t_h}$ and $\rho_{t_g t_h} = 1$

Also, from (7) and (8), we have $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$ for any test.

Then we have

$$\mu_{x_g} = \mu_{x_h} \; and \; \sigma_{x_g}\sigma_{x_h}\,, \qquad \ldots(11)$$

For two parallel tests g and h

Thus the means of raw scores on two parallel tests are equal; and so are the standard deviations.

If we have more than two parallel tests (at least three-say *g, h and k*) .we have another condition to check; besides (11), before we can conclude that the tests *g, h* and *k* are parallel .And this condition is

$$\rho_{x_g x_h} = \rho_{x_g x_k} = \rho_{x_h x_k}\,, \qquad \ldots(12)$$

The condition of equality of all inter-correlations between raw scores of the parallel tests

Now we establish (12) by first obtaining an expression for $\rho_{x_g x_h}$ in terms of $\sigma_i^2 and\ \sigma_x^2$.

$$\rho_{x_g x_h} = \frac{cov(x_g, x_h)}{\sigma_{xg}\ X\ \sigma_{xh}}$$

$$= \frac{cov(t_g, t_h) + cov(t_g, e_h) + cov(t_h, e_g) + cov(e_g, e_h)}{\sigma_{x_g} X \sigma_{x_h}}$$

$= \frac{cov(t_g, t_h)}{\sigma_{x_g}^2}$ (Since g, h are parallel tests, the remaining covariance terms are all zero and $\sigma_{xg} = \sigma_{xh}$).

$= \frac{\rho_{t_g t_h} \sigma_{t_g} \sigma_{t_h}}{\sigma_{x_g}^2}$

$= \sigma_{t_g}^2 / \sigma_{x_g}^2$ (since $\rho_{t_g t_h} = 1$ and $\sigma_{t_g} = \sigma_{t_h}$, g and h being parallel).

Thus for two parallel tests g and h,

$$\rho_{x_g x_h} = \sigma_{t_g}^2 / \sigma_{x_g}^2$$

$$= \sigma_{t_h}^2 / \sigma_{x_h}^2 (since\ \sigma_{t_g} = \sigma_{t_h},\ \sigma_{x_g} = \sigma_{x_h}) \quad \ldots(13)$$

Equ. (13) easily establishes equ. (12) for a number of parallel tests.

Thus for three or more parallel tests the means of raw scores are equal; so are the variances and the inter correlations. In addition to satisfying these criteria, parallel tests should also be similar with respect to the content and nature of items, etc., which may be verified by expert judgment only.

DEFINITION OF TRUE SCORE

Equations (8) define error score. Then the true score (t) can be regarded as the difference $(x - e)$ between the raw score and the error score. Thus, $t_i = x_i - e_i$.

Alternatively, we may define the true score of an individual as the limit of the average of the raw score of the individual on a number of parallel tests k approaches infinity, $i.e$.

$$t_i = \lim_{k \to \infty} \left[\sum_{g=1}^{k} x_{i_g} / k \right] \quad \ldots(14)$$

With this definition of t, the error score is defined as the difference $x - t$; $i.e.,\ \ e = x - t$.

ERROR VARIANCE (STANDARD ERROR OF MEASUREMENT)

From equations (7) and, we have.

$$\sigma_x^2 = \sigma_i^2 + \sigma_e^2,$$

And from equation (13), we have, if g and h are parallel tests,

$$\sigma_t^2 = \rho_{x_g x_h}\sigma_x^2,$$

Thus combining the above two relations we get

$$\sigma_x^2 = \sigma_x^2\rho_{x_g x_h} + \sigma_e^2$$

Or $$\sigma_e^2 = \sigma_x^2(1 - \rho_{x_g x_h})$$

Or $$\sigma_e = \sigma_x\sqrt{1 - }\,\rho_{x_g x_h} \quad \ldots(15)$$

Equation (5.15) gives the standard deviation of the error scores, which is technically known as the ***standard error of measurement.***

DEFINITION OF RELIABILITY

We define reliability as the reproducibility of the measurements when remeasured under identical conditions. Spearman first introduced the term 'reliability'. The reliability of a test (a measuring instrument) is given by the correlation between the raw scores of the given test and a parallel test. Thus, if g be the given test and h any other test parallel to g, then the reliability of g is measured by $\rho_{x_g x_h}$ and will be denoted as ρ_{gg}.

From equation (5.13), we know that

$$\rho_{gg} = \sigma_{t_g}^2/\sigma_{x_g}^2 \quad \Big\}$$

$$= 1 - \sigma_{e_g}^2 / \sigma_{x_g}^2 \qquad \ldots(16)$$

by virtue of the relation $\sigma_t^2 = \sigma_x^2 - \sigma_e^2$.

Reliability can thus be defined as the ratio of the true score variance to the raw score variance or as the proportion of the raw score variance that is the true score variance. Reliability ranges from zero to one. $\rho_{gg} = 1$ when $\sigma_e = 0$. But $\sigma_e = 0$ if and only if all $e_i = 0$, since $\mu_e = 0$. Thus, the test is perfectly reliable ($\rho_{gg} = 1$) if $x_i = t_i$ for each i, and then the raw scores are the true scores. $\rho_{gg} = 0$ if $\sigma_i = 0$ (or , equivalently, if $\sigma_e = \sigma_x$), $i.e.$, when $x_i = t + e_i$ for each i, and then the test is unreliable (here t denotes true score for all i).

For any test g, therefore,

$$0 \leq \rho_{gg} \leq 1.$$

It may be noted , however , that when the reliability is measured from a sample of individuals , one obtain a negative coefficient.

EFFECT OF TEST LENGTH ON THE RELIABILITY OF A TEST

By the length of a test we mean the number of items in the test. Let us augment the length of the test by adding to $(k-1)$ parallel tests of the same length. So the composite test is now made of k parallel test of the same length and the length of the composite test is k times the length of the original test. The effects of this increase in length on the true score variance and raw score variances are the following:

Denoting the k parallel tests by g_1, g_2 ..... g_k and the composite test by G, we have

$$\sigma_{i_G}^2 = \sigma^2\left(t_{g_1} + t_{g_2} + \cdots + t_{g_k}\right) = \sum_i \sum_j \rho t_{g_i} t_{g_j} \sigma_{t_{g_i}} \sigma_{t_{g_j}}$$

(Summation over $i, j = 1, 2, \ldots \ldots, k$)

$= k^2\sigma_{t_{g_1}}^2$ (since the component tests are parallel,

$$\rho_{t_{g_i}t_{g_j}} = 1 \text{ and } \sigma_{t_{g_i}} = \sigma_{t_{g_j}} \quad (\text{for all } i, j) \qquad \ldots(17)$$

and, $\sigma_{x_G}^2 = \sigma^2\left(x_{g_1} + x_{g_2} + \cdots + x_{g_k}\right) = \sum_{i=1}^{k} \sigma_{x_{g_i}}^2 + \sum_{i}\sum_{\neq j} \rho_{x_{g_i}} x_{g_i}\sigma_{x_{g_i}}\sigma_{x_{g_j}}$

$$= k\sigma_{x_{g_1}}^2 + k(k-1)\rho_{gg}\sigma_{x_{g_1}}^2 \qquad \ldots(18)$$

Since $\rho_{x_{g_i}x_{g_j}} = \rho_{gg}(i.e\ reliability)$ and $\sigma_{x_{g_i}} = \sigma_{x_{g_j}}$ for parallel tests g_i, g_j

Using equation (16), we may write down the reliability of a test whose length is increased k times (by adding $k-1$ parallel tests) as

$$\rho_{GG} = \sigma_{i_G}^2/\sigma_{x_{G,}}^2$$

which can be expressed in terms of ρ_{gg}, by using equation (15) and (18) as,

$$\rho_{GG} = \frac{k^2\sigma_{i_{g_1}}^2}{k\sigma_{x_{g_1}}^2[1 + (k-1)\rho_{gg}]}$$

$$= \frac{k\rho_{gg}}{1+(k-1)\rho_{gg}} \qquad \ldots(19)$$

where ρ_{gg} is the reliability of the original is test and ρ_{GG} is the reliability of the lengthened test G, whose length is equal to k times the length of g_1.

Formula (19) is known as the general ***Spearman brown formula***. In the usual case where $k = 2$, Spearman–Brown formula or doubled test length is

$$\rho_{GG} = \frac{2\rho_{gg}}{1+\rho_{gg}} \qquad \ldots(20)$$

The derivation of formula (19) and (20) involves the assumption that the additional test parts used in lengthening the original test are parallel to those in the original test

The formula for determining k is obtained by solving equation (19) for k:

$$k = \frac{\rho_{GG}(1-\rho_{gg})}{\rho_{gg}(1-\rho_{GG})}, \qquad \ldots(21)$$

Where $\rho$the reliability of the original test and ρ_{GG} is is the desired reliability of the lengthened test after the original test is lengthened k

Example. What would be the reliability coefficient when the original test of reliability 0.50 would be doubled in length?

We have in this case $\rho_{gg} = 0.50$ and $k = 2$. then by equation (20) we get, as the reliability of the lengthened test

$$\rho_{GG} = \frac{2 \text{ X } .50}{1 + .50} = 0.67.$$

Example. By what amount should the length of a test of reliability 0.66 be increased so as to get a reliability of 0.95 for the lengthened test?

Here $\rho_{gg} = 0.67$ and $\rho_{GG} = 0.95$. Then by equation (21), we have

$$k = \frac{.95(1-.67)}{.67(1-.95)} = \frac{.95 \text{ X } .33}{.67 \text{ X } .05} = \frac{.3135}{.0335} = 9(\text{Approximately}).$$

PRACTICAL METHOD OF ESTIMATING TEST RELIABILITY

Reliability, as defined above and denoted by ρ_{gg}, is based on population data (an infinite number of individuals being tested). In practice, we have only a sample of finite size n and the corresponding sample correlation estimated the reliability. There are available mainly four methods for estimated test reliability. These are:

(a) The parallel–test method, (b) the test-retest method, (c) the split–half method and (d) the Kuder-Richardson method.

Parallel –test method

Reliability was defined as the correlation between raw scores on two parallel tests. In this method, two tests are constructed satisfying as far as possible the conditions for parallelism. Then the two tests are administered to

the same group with a suitable time lag and the reliability (ρ_{gg}) is estimated by the correlation(r_{gg}) between the raw scores of the parallel test obtained from the sample.

For many situations, this is the best method of estimating test reliability. However, the ability measured should not change in the time interval between the administrations of the test. For many scholastic achievement and mental ability tests, this condition is fulfilled. But there are case where the ability tested will change, $e.g.$, in performance tests like type-writing tests, athletic skills tests etc., if the individuals continue practicing during the interval between the two administrations.

The parallel- test reliability may also be obtained by administering both the tests at the same session , In this case also the scores on the second test may be influenced either by familiarity with the material in the first test or by fatigue.

Generally speaking, parallel –test reliability will give a satisfactory result .But the difficulty is to construct two parallel test. So when only one test is available, we are to use one of the other methods.

Test-retest method

This method consists in administering the same test twice after a suitable time interval to eliminate familiarity with the material, test fatigue, etc., and then finding the correlation between the test scores and retest scores. If, however, the individuals duplicate their first performance, then the reliability will be over-estimated by this method.

If the test is repeated immediately, the memory effect, practice and confidence will increase the scores on retesting. If sufficient time elapses before the second administration, then these effects will be absent and the test-retest correlation will give an estimate of the stability of the test scores.

As in the parallel-test method, here also, the experimenter will have to adjust the time interval and control the activity of the individuals within the time interval so as to minimize the effects due to memory, fatigue practice etc.

The difficulty with both these methods is that sometimes it is difficult to get the individuals again after an interval of time. In such a case, we cannot apply either the same test twice or two parallel–tests. For such case, we have the following methods.

Split-half method

Here one test is applied once and then the score is divided into two equivalent halves, and the correlation between the score on the half-tests estimates the reliability of each half-tests. Then by Spearman- Brown formula (5.20), we may estimate the reliability of the original (full) test.

The test may be split into two parts in a number of ways. The commonest way is to split the test on the basis of odd-numbered and even-numbered items.

In many performance tests or personality tests, it is difficult to construct parallel test or to retest with the same test. So the split-half method is regarded as the best method in such cases. The objection that is often raised is that there is no unique way of splitting the test and unique split-half correlation. In most *Power test* (where one dose not emphasize the speed or quickness with which the work can be performed), the items are arranged in order of difficulty, and the odd-even split provides a unique estimate of reliability

Rulon presented the following formula for estimating reliability from two subtest scores (of the same test):

$$r_{gg} = 1 - \frac{s_d^2}{s_x^2} \qquad \ldots(22)$$

where s_x^2 is the variance of raw scores and s_d^2 is the variance of the difference of raw scores on the two halves of the test.

Similar results may be obtained by using the formula due to Guttman, which is similar to apply:

$$r_{gg} = 2\left[1 - \frac{s_1^2 + s_2^2}{s_x^2}\right], \qquad \ldots(23)$$

where s_1^2 and s_2^2 are the variances of raw scores on the two halves.

Equations (20), (22) and (23) will give the same reliability coefficient when $s_1^2 = s_2^2$, $i.e.$, when the two halves have equal raw scores variances. If $s_1^2 \neq s_2^2$, then the split-half reliability given by equ.(20) will be the highest.

Kuder-Richardson method

We shall obtain the Kuder-Richardson formulae for estimating test reliability by making the same assumptions as were made originally by Kuder and Richardson. Let us consider a test of length k which is made up of k parallel items .Then the raw score variance is given by

$$\sigma_x^2 = \sigma^2_{(x_1+x_2+\cdots+x_k)} = \sum_{g=1}^{k} \sigma_{x_g}^2 + \sum_{g} \sum_{\neq h} \rho_{x_g x_h}\, \sigma_{x_g} \sigma_{x_h}.$$

Since the items are all parallel $\rho_{x_g x_h}$ will be equal to ρ_{gg} (reliability of item g) for all g and h and σ_{x_g} will be the same for all g. Thus,

$$\sigma_x^2 = k\sigma_x^2 + k(k-1)\, \rho_{gg} \sigma_{x_g}^2,$$

so that the item reliability (ρ_{gg}) can be expressed as follows:

$$\rho_{gg} = \frac{\sigma_x^2 - \sum_{g=1}^{k} \sigma_{x_g}^2}{(k-1) \sum_{g=1}^{k} \sigma_{x_g}^2}, \qquad since \sum_{g=1}^{g} \sigma_{x_g}^2 = k\sigma_{x_g}^2$$

Next, to obtain the reliability of the test of k parallel items from ρ_{gg}, we apply the general Spearman–Brown formula (19):

$$\rho_{GG} = \frac{k\rho_{gg}}{1 + (k-1)\rho_{gg}}$$

$$= k \frac{\sigma_x^2 - \sum_{x=1}^{k} \sigma_{x_g}^2}{(k-1) \sum_{x=1}^{k} \sigma_{x_g}^2} \text{ x } \frac{1}{1 + (k-1)[(\sigma_x^2 - \sum_{x=1}^{k} \sigma_{x_g}^2)/(k-1) \sum_{x=1}^{k} \sigma_{x_g}^2]}$$

$$= \left[\frac{k}{k-1}\right] \text{x} \left[\frac{\sigma_x^2 - \sum_{x=1}^{k} \sigma_{x_g}}{\sigma_x^2}\right]. \qquad \ldots(24)$$

This is the Kuder-Richardson "formula 20" for obtaining the reliability of a test of k parallel items in terms of k, σ_x^2 and $\sigma_{x_g}^2$. In practice, this is estimated by

$$r_{GG} = \left[\frac{k}{k-1}\right]\left[\frac{s_x^2 - \sum_{g=1}^{k} s_{x_g}^2}{s_x^2}\right] \qquad \ldots(24a)$$

where s_x^2 is the sample variance of raw total scores and $s_{x_g}^2$ is the same for g.

If the scoring of items be 1 for a correct response and 0 for wrong response, then $s_{x_g}^2 = p_g(1 - p_g)$, where p_g is the sample proportion of correct response for item g. Then formula (24a) simplifies to

$$r_{GG} = \left[\frac{k}{k-1}\right]\left[\frac{s_x^2 - \sum_{g=1}^{k} p_g(1-p_g)}{s_x^2}\right] \qquad \ldots(25)$$

If in formula (24) we assume that the k parallel items are of equal difficulty, the scorning being 1 for a correct and 0 for a wrong response, with π as the common difficulty value for all items, then

$$\sigma_{x_g}^2 = \pi(1 - \pi) = \pi - \pi^2$$

Now, the mean of obtained scores on the test is

$$\mu_x = k\pi$$

Thus,

$$\sigma_{x_g}^2 = \frac{\mu_x}{k} - \frac{\mu_x^2}{k^2},$$

Then from formula (24), we have

$$\rho_{GG} = \left[\frac{k}{k-1}\right]\left[1 - \frac{k\sigma_{x_g}^2}{\sigma_x^2}\right]$$

$$= \left[\frac{k}{k-1}\right]\left[1 - \frac{\mu_x - \mu_x^2/k}{\sigma_x^2}\right] \qquad \ldots(26)$$

This is the Kuder-Richardson "formula 21" for obtaining the reliability of a test of k parallel items of equal difficulty in terms of k, σ_x^2 and μ_x. In practice this is estimated by

$$r_{GG} = \left[\frac{k}{k-1}\right]\left[1 - \frac{\bar{x} - \bar{x}^2/k}{s_x^2}\right] \qquad \ldots(26a)$$

where $\bar{x}$ and s_x^2 are the sample mean and variance of raw total scores.

We have divided the Kuder-Richardson formula under original assumptions. However it is also possible to derive them under less restrictive conditions.

VALIDITY

In the previous section, we considered one essential property of a measuring instrument – the ***reliability.*** Now we shall consider the second essential property– the validity. A psychological test (a measuring instrument) should not only be reliable, but it should also be valid. By this we mean that the test should measure what it is supposed or intended to measure. If we want to measure a trait A for a group of individuals with the test, we must be sure, before we can use the test confidently for that purpose, that it actually measures the trait A and also measures it reliably. The term 'validity' is a relative term-a test is valid for a particular trait for a particular group or for a particular situation. We may use the same test for measuring different traits and then we must obtain the validity separately for each case.

As with the reliability of physical measurements, in the case of the validity of such measurements also, we face no great problem. But the situation is different with psychological measurements.

To estimate the validity of a test we must know which particular trait we want to measure. We make use of some known measure of the trait called the ***criterion variable***. The validity of the test is then estimated by computing a coefficient (the ***coefficient of validity***) which determines relationship between the scores obtained on the test and the values of the criterion

variable and getting measures on this variable which are to be compared with the scores on the test. Often it is difficult to get reliable measures on the true criterion. What we get are only approximate measures on the criterion variable. Depending upon the situation, the criterion scores may be of any of the following kinds: ratings by judges (experts who know the group) on the trait measured scores on another valid test of the (we may validate a newly constructed test for trait A by selecting as the criterion variable the score on a well–established test for trait *A*), measure of later success (for a test for recruiting persons in a vocation), etc. We discuss below the different concepts of validity:

Predictive validity

This type of validity arises when we use a test for trait for selecting applicant for a particular course or job and the criterion variable is the degree of success at a later period, *i*. *e*., after the recruits have completed the course or have been on the job for a sufficient period. The criterion variable is the performance at that later period– grades or ratings on completion of the course or after a certain period of employment. A test has a high predictive validity if it can forecast efficiently later performance on a particular measurable aspect of life. And this is of importance in the selection or recruitment of individuals for different courses of study or training programmes or jobs.

Concurrent validity

Concurrent validity is obtained for tests for which the criterion variable is also available at the same times as the test results and we are not to wait as in the case of predictive validity. Tests are constructed for measuring a variable for which the result also may be obtained without waiting, because it is easier and sometimes saves time and expenditure, while giving the same results as the criterion variable. Concurrent validity is used for diagnostic test (*e*. *g*. in clinical diagnosis). Both types of validity (predictive and concurrent)

are obtained by computing the correlation between the test scores and criterion scores and the validity is the correlation coefficient.

Content validity

Sometimes tests are constructed to study the knowledge of the individuals on certain specific areas of study, say verbal ability, geometrical drawing ability, etc. There are large numbers of items which measure these areas and, in a test, we have only a sample of these items. In content validity of a test, we try to ascertain how far the test covers the field of study under investigation or in other words, how good the items of the test are as a sample from the totality of all items for that test. It is, however, not possible to express content validity as a validity coefficient, as is possible with the previous two validities.

Construct validity

This is comparatively a new concept in validity theory. This concept is found useful when either there is no external criterion variable or it is difficult to obtain measurements on the criterion variables. This validity cannot be expressed in a single measure as the correlation between test scores and criterion scores. Validity in this case is demonstrated by showing that the predictions expected on the basis of theory may be confirmed by the test. Some of the common ways of establishing construct validity are the following:
(1) Correlating different items or parts of the test. There correlations should be high if the test is measuring a unitary variable.
(2) Correlating different tests which measure the same variable.

CORRECTIONS FOR ATTENUATION:

A validity coefficient expresses the extent of agreement of the test score with a measurement of the criterion variable. Both these measurements are, however, liable to errors, which are due to unreliability of the measuring instruments. It is possible to develop a correlation for these errors, known as the correction for attenuation.

The corrected value of the validity coefficient will estimate the relationship of the test score and the criterion score, had both the measurements been completely reliable.

Let T_i and C_i be the observed test score and criterion score for the ith individual, t_i and c_i the corresponding true scores, and e_i and e_i' the errors. Thus,

$$T_i = t_i + e_i \text{ and } c_i = e_i + e_i'$$

all expressed as deviations from means.

Thus r_{ic} the true validity coefficient, is

$$r_{tc} = \frac{\sum(T_i - e_i)(C_i - e_i')}{N s_t s_c}$$ (N being the total number of individuals),

so that

$$r_{tc} s_t s_c = \frac{\sum T_i C_i}{N} - \frac{\sum T_i e_i'}{N} - \frac{\sum C_i e_i}{N} + \frac{\sum e_i e_i'}{N}$$

Assuming independence of true scores and error scores and of error scores themselves,

$$r_{tc} = \frac{r_{TC} s_T s_C}{s_t s_c}$$

From (5.16), we know

$$r_{TT} = \frac{s_t^2}{s_T^2} \text{ and } r_{CC} = \frac{s_{C^2}}{s_{C^2}}$$

r_{TT} and r_{CC} being estimates of reliability of test score and criterion scores.

Thus,

$$r_{tc} = \frac{r_{TC}}{\sqrt{r_{TT} r_{CC}}} \quad \ldots(27)$$

But this coefficient is of little practical value, since a pair of perfectly reliable test and criterion is rarely realized. Very often we shall be using test scores which are contaminated with errors for the purpose of prediction. There, it

may be of interest to know what would be the validity coefficient had a perfectly reliable criterion been available. In the same way, we can find the correlation between true criterion score and observed test score, as

$$r_{tc} = \frac{r_{TC}}{\sqrt{r_{CC}}}. \qquad \ldots(28)$$

EFFECT OF TEST LENGTH ON TEST PARAMETERS

We have seen in previous section, the effect of test length on the true score variance (equ.17), on the observed score variance (equ.18) and on the reliability of a test (equ.19). Using notations already introduced, it is easy to see the effect of test length on true mean and observed score mean:

$$\mu_{t_G} = k\mu_{t_g} \qquad \ldots(29)$$

and

$$\mu_{x_G} = k\mu_{x_g} \qquad \ldots(30)$$

To find the effect of test length on the validity of a test, we first consider the case where the original test is lengthened by adding to it $(k-1)$ parallel test of the same length and the original criterion variable is lengthened by adding to it $(l-1)$ parallel criterion variables of the same length, such that each pair of component test and criterion variable gives the same validity coefficient.

Let us denote the total test score by x_G :

$$x_G = x_{g_1} + x_{g_2} + \cdots \ldots \ldots + x_{g_k}$$

and the total criterion score by y_H:

$$y_H = y_{h_1} + y_{h_2} + \cdots \ldots . + y_{h_i} .$$

Now we obtain the correlation coefficient of augmented test scores with the augmented criterion variable scores:

$$\rho_{x_G y_H} = \frac{cov(x_G, y_H)}{\sigma_{x_G} \text{ X } \sigma_{y_H}}$$

$$= \frac{cov(x_{g_1} + x_{g_2} + \cdots + x_{g_k}, y_{h_1} + y_{h_2} + \cdots + y_{h_l})}{\sqrt{var(x_{g_1} + x_{g_2} + \cdots + x_{g_k}) \text{ X } var(y_{h_1} + y_{h_2} + \cdots + y_{h_l})}}$$

$$= \frac{\sum_{g=1}^{k} \sum_{h=1}^{l} \rho_{x_g y_h} \sigma_{x_g} \sigma_{y_h}}{\left\{k\sigma_{x_g}^2 + k(k-1)\rho_{gg}\sigma_{x_g}^2\right\}^{1/2} \{l\sigma_{y_h}^2 + l(l-1)\rho_{hh}\sigma_{y_h}^2\}^{1/2}}$$

$$= \frac{kl\rho_{x_g y_h}\sigma_{x_g}\sigma_{y_h}}{\left\{k + k(k-1)\rho_{gg}\right\}^{1/2}\{l + l(l-1)\rho_{gg}\}^{1/2}\sigma_{x_g}\sigma_{y_h}}$$

$$= \frac{kl\rho_{x_g y_h}}{\left\{k + k(k-1)\rho_{gg}\right\}^{1/2}\{l + l(l-1)\rho_{hh}\}^{1/2}}$$

... (31)

where $\rho_{x_g y_h}$ is the validity of the original test with the original criterion variable.

$\rho_{x_G y_H}$ is the validity of the lengthened test(lengthened k times),

with the lengthened criterion variable (lengthened l times),

ρ_{gg} is the reliability of the original test and

ρ_{hh} is the reliability of the original criterion variable.

If the criterion variable is not lengthened, then the effect on the validity of increasing only the test length is obtained from (5.31) by putting$l = 1$:

$$\rho_{x_G y_H} = \frac{k\rho_{x_g y_h}}{\{k+k(k-1)\rho_{gg}\}1/2} \quad \text{...(32)}$$

ITEM ANALYSIS

We have already seen that in constructing a test will be determined by its reliability and validity. Now, in developing a test a large number of items supposed to measure the ability under consideration are tried over a large group of subjects. The question that naturally arises is : how well can the item be selected so that the required reliability and validity of the test can be achieved? This calls for item analysis

The typical item analysis is carried out from two kinds of information – an index of item difficulty and an index of item validity, which means how well the item discriminates in agreement with the rest of the items of the test or how well it predicts some external criterion. The most common index of item difficulty is p_i, the proportion of subjects who pass the item. The commonly used index of item validity is r_{ie}, the correlation of the item score with some external criterion c or, more often r_{ic}, the correlation of the item score with the total score. The most common use of item analysis data is the selection of the best items to compose the final test. It also enables the item-writer to modify the items in the required directions. The important features of the test, viz. mean, variance, reliability and validity, can be controlled by selecting items of the right type of difficulty, the right spread of difficulty, the right degree of item inter correlations and item validities.

The difficulty index p_i for the ith item is the proportion of subjects answering the item correctly. In a multiple–choice item with k alternatives, Guilford has proposed a ***correction for guessing*** on the assumption that a subject either knows the answer correctly or guesses at random. If R_i is the number of persons answering the item correctly & W_i the number answering wrongly, the number of lucky guesses, $i.e$ of those who guess correctly, is estimated as $\frac{W_i}{k-1}$ so that the item difficulty corrected for guessing is

$$\frac{R_i - \frac{W_i}{k-1}}{R_i + W_i} \qquad \ldots(33)$$

There are alternative formulae for correction for guessing too, based on other assumptions. In some methods of item analysis, the correction r_{it} is estimated

from those making extreme scores, generally the upper & lower 27% of total group. The estimation is, however, based on symmetry of the item score & total score distributions & linearity of regression of item score on total score.

Four coefficients of correlation are commonly used to indicate the correlation of an item with a criterion (r_{ic}) or, more generally, of an item with the total(r_{it}). They are ***biserial***($\boldsymbol{r_{bi}}$)***, point biserial*** $\left(\boldsymbol{r_{pb}}\right)$, ***tetrachoric*** ($\boldsymbol{r_t}$) ***and the Φ coefficient.*** If the ability measured by the item is normally distributed and the criterion score is continuous, then r_{bi} can be used. If the item score is limited to 0 and 1, r_{pb} should be used. If the criterion variable and the ability measured by the item are both normally distributed, r_t is called for. If the criterion is not a continuous variable, but a natural division into two groups, one can use the Φ coefficient. Another index, known as the index of discrimination between High and Low groups, is often used for item selection.

INTELLIGENCE TESTS AND IQ

Interest in the nature and measurement of intelligence is gradually increasing. Tests of intelligence and other mental qualities are being used in different spheres of life. By intelligence is meant the capacity for relational and constructive thinking for the attainment of some goal. In the discussion of intelligence, Spearman's two–factor theory holds an important place. According to this theory, there is a common element, a ***general factor***, in all our **cognitive abilities**- abilities that are concerned with the intellectual aspects of mind. Spearman named this as the ***g*-**factor and this ***g*-**factor can be identified with intelligence. Besides the ***g*-**factor, which is present in all abilities, there is according to Spearman a ***specific factor*** for each ability. Spearman's theory was not, however, universally accepted. Thomson proposed a group- factor theory. According to Thomson, there are ***group factors*,** each of which is present in a number of different abilities. Thus, while they are more restricted than Spearman's ***g*-**factor, they are less restricted than his specific factors. Some of the group factors are the following (i) verbal ability; (ii) numerical ability; (iii) musical ability; (iv) mechanical ability;

All attempts to describe intelligence by a recourse to physiology have failed. Though differences of opinion exist on nature of intelligence, there is more or less general agreement as to the procedure of measuring intelligence. In an intelligence test, the following types of problem find a place:

(i) Synonyms and antonyms

One word is given, and the subject is required to select or to supply a second word which has the same or the opposite meaning.

Example: (i) Superior is the opposite of …….. .

(ii) Cruel is the same as (rough, unkind, persecutor, inhuman).

(ii) Classification

A set of word is given. All but one word are in some respect the same. The subject is to find out the odd word.

Example: (i) Shoot, stab, murder, write.

(ii) Rice, flour, bread, flower.

(iii) Sentence completions

An incomplete sentence is given. The subject is to complete it.

Example: (i) Man is superior to other animals because…….,

(ii) A journey to moon can be made by……..,

(iv)Mixed sentences

A set of words is given. The subject is to rearrange them into a sentence and say whether it is true or false.

Example: (i) Sword pen is then mightier. (True, False)

(ii) Is America a socialist country.(true, false)

(v)Coding

A sentence is given. The subject is to rewrite it on the basis of a given code.

Example: Code the following message by first reversing each word and then

substituting each letter by the next- “Send reinforcements at once”.

(vi)Number series

A series of numbers is given and the subject is to supply the next or the next two.

Example: (i) Supply the next two terms-.

(a) 1, 3, 7, 13, ….,…..,… .

(b) 81, 27,9, 3….,….,… .

(vii) Analogies

Three words, of which the first two are related in some way, are given. The subject is to find or select the fourth word which is related to the third as the second is to the first.

Example: Black is to white as intelligent is to ……. .

Man is to woman as god is to ……. .

(viii) Inferences

A problem demanding reasoning is given, and the subject is to select or supply the solution.

Example: All men are mortal.

Some men are kind.

All mortals are kind. (True or false)

Intelligence tests may be designed for application to individuals or for application to groups of individuals. One of the well-known individual tests is ***Binet’s test.*** The revised version of this test is now being widely used for

measuring Intelligence of young children and for detecting mental deficiency. Group tests were first widely used by the U.S Army authorities for recruitment, placement or promotion of personnel. The Alpha test was meant for the majority and the Beta test for illiterates or non-English-Speaking person.

Intelligence tests, like other tests, may again be verbal or non-verbal, The former demand the intelligent manipulation of ideas expressed in words while the letter call for the intelligent manipulation of objects.

After constructing an Intelligence test, we must check its reliability and validity by one of the methods discussed previously. When we are satisfied that the Intelligence test is reliable and valid, we must compute some standard or *norm* which will aid us in assessing any given individual's score. We may compute either the mean and standard deviation or the percentile norms, standard scores or T–sores for this purpose. It was in this connection that Binet introduced the concept of ***mental age.*** An individual's mental age (*MA*) is the age at which an average person can pass the tests that the individual passes. A numbers of intelligent tests so constructed are to be applied to large numbers of children of different ages. Then one has to find at what age last birth day each test is passed by 50% of the children of that age. Thus for each age a number of intelligence tests, say 5, are fixed. If a subject can answer correctly all the tests for age 9, 80% of age 10, 40% of age 11 and 20% of age 12, his mental age would be 9+.80+40+20=10.40. Later, ***mental ratio (MR)*** was defined as

$$metal\ ratio = \frac{mental\ age}{chronological\ age} \qquad \ldots(34)$$

Thus, if a boy of 10 years possesses an *MA* of 10.40 years, then his *MR* is 1.04. He is thus an advanced child, his *MR* being more than 1. A child will be regarded as retarded if his *MR* is less than I, and he is of average intelligence if his *MR* equals I.

The ***intelligence quotient, or IQ,***has now replaced the *MR*. IQ is defined as

$$IQ = 100 \text{ X } \frac{MA}{CA}$$

$$= 100 \text{ X } MR \qquad \ldots(35)$$

We now make some observations concerning the interpretation of IQ in its classical from. The IQ will be 100(lower than100/greater than 100) for all children who have the same (a lower/a higher) level of intellectual development as (than) the average child of the same age. It is necessary that the standard deviations of the IQ distribution of all age groups be approximately the same for the same IQ to have the same relative position on the distribution for different ages. This is essential for a proper interpretation of an individual IQ. But as this is not fulfilled in many cases, the present trend in standard tests is that the test is standardized and normalized into a set of normalized scores (called IQ –equivalents) for each age with mean 100 and standard deviation 15. Thus it is immaterial whether we use a $\mathcal{I}$-scale or an IQ–equivalent scale for the norm.

The use of intelligence tests has shown that intelligence may be supposed to be normally distributed and that it depends on heredity. It has also been found that intelligence grows with age, which continues up to age 16 or 17, and then it remains steady. There is no evidence that intelligence and sex are related. It has also been found that different occupations require intelligence to varying degrees.

Intelligence tests have found many uses. They are used for vocational guidance and selection, in the grading of pupils and in diagnosing mental deficiency. Thus an intelligence test, properly constructed and standardized, is of immense use for various purposes.

ELEMENTS OF FACTOR ANALYSIS

Factor analysis is that branch of statistical methods which is concerned with the resolution of a set of variables X_1, X_2,……..Xn in terms of a smaller number of factors F_1, F_2…….. F_m, where $m < n$ so that the purpose in views

is not vitiated. The resolution is effected by the analysis of inter-correlations of the variables. The satisfactory solution is to use factors which convey all the important and essential information of the original set of variables and the emphasis is on economy of description. Factor analysis has its principal application in psychological measurements, where the variables X_1, X_2,……..Xn are the test scores on n score of a battery and F_1, F_2,……,F_m are m mental abilities measured by the tests.

The simplest mathematical expression for describing a set of variables in terms of several others is a linear one. In factor analysis also, a linear form is taken to represent a variable Xj in terms of a number of underlying factors which are taken in the standardized form ($i.e.$, with zero means and unit $s.d.$ ''''''''''''s). Several types of factors are employed. Common factors are those which occur in more than one variable. Common factors are of two types - (1) general factor, which is common to all the variables and (2) group factors, which are present in several, but not in all, variables. A factor which appears in the description of a single variable is called ***unique.*** Unique factors are of two types – (1) specific factors, having a simple interpretation and liable to be identified, and (2) unreliable or error factors, which are unreliable and not identifiable. Thus we have

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \cdots + a_{jm}F_m + b_jS_j + c_jE_{J,} \qquad j = 1,2 \ldots . n, \qquad \ldots(36)$$

F_1, F_2, ……, F_m being the common factors S_j the specific factor and $E_{j,}$ the error or unreliability.

$h_j^2 = \sum_{k=1}^{m} a_j^2 k$ is called the ***communality*** of the variable $X_{j,}$ which is the part of the total variance attributable to common factors, whereas b_j^2 and c_j^2 are called the ***specificity*** and ***unreliability*** of the variable, $b_j^2 + c_j^2$ being called its ***uniqueness***. $h_j^2 + b_j^2$ may be termed as the ***reliability*** of the variable, and $a_{j1,}a_{j2,}$……,$a_{jm,}$ are the ***factor loadings*** of the m common factors for the variable X_j. The basic problem of factor analysis is to determine the factor loadings. When the factor loadings are determined one can evaluate the factors in terms of the variables.

Let us designate

$$X_j = a_{i1}F_1 + a_{j2}F_2 + \cdots + a_{jm}F_m + a_jU_j, \quad j = 1,2 \ldots . . n, \qquad \ldots(37)$$

U_j Being the uniqueness, as the factor pattern and

$$\left.\begin{aligned} r_{X_jF_k} &= a_{j1}r_{F_kF_1} + a_{j2}r_{F_kF_2} + \cdots + a_{jk} + \cdots + a_{jm}r_{F_kF_m} \\ r_{x_j}U_j &= a_j \end{aligned}\right\} \quad \ldots(38)$$

as the factor structure

If we have N individuals for whom the values of the variable X_j are known, say $X_{j1}, X_{j2}, \ldots\ldots, X_{jN}$,let

$$X = \begin{pmatrix} X_{11} & X_{12} \ldots \ldots X_{1N} \\ X_{21} & X_{22} \ldots \ldots X_{2N} \\ \ldots & \ldots \quad \ldots \ldots \quad \ldots \\ X_{n1} & X_{n2} \ldots \ldots X_{nN} \end{pmatrix}$$

$$F = \begin{pmatrix} F_{11} & F_{12} \ldots \ldots & F_{1N} \\ F_{21} & F_{22} \ldots \ldots & F_{2N} \\ \ldots & \ldots \quad \ldots \ldots & \ldots \\ F_{m1} & F_{m2} \ldots \ldots & F_{mN} \\ U_{11} & U_{12} \ldots \ldots & U_{1N} \\ \ldots & \ldots \quad \ldots \ldots & \ldots \\ U_{21} & U_{22} \ldots \ldots & U_{2N} \\ U_{n1} & U_{n2} \ldots \ldots & U_{nN} \end{pmatrix}$$

And

$$M = \begin{pmatrix} a_{11} & a_{12} \ldots \ldots a_{1m} & a_1 & 0 & 0 \ldots \ldots 0 \\ a_{21} & a_{22} \ldots \ldots . a_{2m} & 0 & a_2 & 0 \ldots \ldots 0 \\ \ldots & \ldots \quad \ldots \ldots \quad \ldots & \ldots & \ldots & \ldots \quad \ldots \ldots \quad \ldots \\ a_{n1} & a_{n2} \ldots \ldots . a_{nm} & 0 & 0 & 0 \ldots \ldots a_n \end{pmatrix}$$

Then

$$X = MF.$$

Now

$$\frac{1}{N}XX = \begin{pmatrix} 1 & r_{12} \ldots \ldots & r_{1n} \\ r_{21} & 1 \ldots \ldots . & r_{2n} \\ \ldots & \ldots \ldots \ldots & \ldots \\ r_{n1} & r_{n2} \ldots \ldots & r_{nn} \end{pmatrix} = R,$$

the correlation matrix

Thus
$$R = \frac{1}{N}XX'$$
$$= \frac{1}{N}(MF)(F'M')$$
$$= M\left(\frac{1}{N}FF'\right)M'$$

But if the factors are all orthogonal,

$$R = MM'.$$

Thus, if we regard the correlation matrix R as the available data and the factor pattern matrix M as the desired objective in a factor analysis, we have $\frac{n(n-1)}{2}$ experimentally given coefficients which must exceed the number of linearly independents coefficients in M. It will be seen that by limiting ourselves to common factors, the factor problem becomes determinate even though we admit the existence of unique factors.

Now with the assumption of a particular factor pattern and the assumption of orthogonality of factors, we can calculate the coefficients

$$\hat{r}_{jk} = \sum_{i=1}^{m} a_{ji}a_{ki}$$

and compare them with the observed correlation coefficients to see how far the assumed factor pattern explains the observed correlation coefficients.

When the factor loadings are determined, estimation of any common factor F_s (or an unique factor U_s) involves the determination of the regression function

$$\hat{F}_s = \beta_{s1}X_1 + \beta_{s2}X_2 + \cdots + \beta_{sn}X_n.$$

The normal equations will be

$$\beta_{s1} + r_{12}\beta_{s2} + \cdots + r_{1n}\beta_{sn} = t_{1s,}$$
$$r_{21}\beta_{s1} + \beta_{s2} + \cdots r_{2n}\beta_{sn} = t_{2s}$$
$$r_{n1}\beta_{s1} + r_{n1}\beta_{s2} + \cdots \beta_{sn} = t_{ns},$$

where $t_{js} = r_{X_j F_s}$

The solution is

$$\hat{\beta}_{sj} = \frac{1}{R}\left[t_{1s}R_{1j} + t_{2s}R_{2j} + \cdots t_{ns}R_{nj}\right],$$

where R_{ij} is the cofactor of r_{ij} in the determinant R=[R].

Thus $\hat{\beta}'_s = t'_s R^{-1}$

so that $\hat{F}_s = t'_s R^{-1}(X_1, X_2, \ldots\ldots, X_n)$ '

Combining for all factors, common and unique, we have

$$F = S'R^{-1}X, \qquad \ldots(39)$$

where $S = \begin{pmatrix} t_{11} & t_{12} & t_{1m} & a_1 & 0 \ldots\ldots 0 \\ t_{21} & t_{22} & t_{2m} & 0 & a_2 \ldots\ldots 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \ldots \ldots \ldots \\ t_{n1} & t_{n2} & t_{nm} & 0 & 0 \ldots\ldots a_n \end{pmatrix}$

In case the factors are orthogonal,

$$r_{X_j F_k} = t_{jk} = a_{jk}$$

and the factor structure coincides with the loading matrix M.

Where $M = \begin{pmatrix} a_{11} & a_{12} & a_{1m} & a_1 & 0 \ldots\ldots 0 \\ a_{21} & a_{22} & a_{2m} & 0 & a_2 \ldots\ldots 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \ldots \ldots \ldots \\ a_{n1} & a_{n2} & a_{nm} & 0 & 0 \ldots\ldots a_n \end{pmatrix}$

We have $\hat{F} = M'R^{-1}X$ $\qquad \ldots(40)$

In actual applications, the orthogonal factors are estimated conveniently by the method of pivotal condensation.