RELIABILITY OF TEST SCORES

A test is like a measuring instrument. One of the important characteristics of any instrument or evaluation device is how reliably it measures. In the simplest of the non-technical language, *reliability means* consistency. If the instrument is reliable, it should give consistent results. In other words, a reliable instrument will give *trustworthy* and *stable results* if it is applied to the same individuals or object from time to time, provided the trait being measured has not itself changed in the meantime. Similarly a reliable test is one which, when applied to same subjects (persons) on different occasions, yields stable and trustworthy results, relatively free from the errors of measurement. For example, if an individual in an intelligence test obtains a raw score of 90, we would expect to find that if we retest him about 2-3 weeks later with the same or parallel test, his score is more or less near 90. On the other hand scores made on unreliable test are subject to larger errors of measurement and are neither stable nor trustworthy. An unreliable test, on repetition, will give inconsistent results.

In modern test theory, "every obtained score is thought of as being made up of two parts, a component which is called the *true score* and a second component called the *error score*. Symbolically, modern test theory can be expressed by the following *linear model.*

$$X_t = X_\infty + X_e$$

where X_t = Obtained or Raw score or measure

X_∞ = True score or measure

$X_e = X_t - X_\infty$, is the Error score or measure

A number of assumptions are made in the model (8·11).

1. The true score (X_∞) is assumed to be the genuine value of the trait being measured, the value we expect on using a perfect instrument under ideal conditions. it may also be defined as the mean of a very large number of determinations made of the given person on the same test or parallel forms of test given under approximately identical conditions. Both the interpretations are consistent. A true score cannot, of course, be determined experimentally.

2. The error component (X_e) of the score is that part which is attributed to such factors as temporary characteristics of an individual, *viz.*, health, fatigue, emotional upset, differences in motivation, etc., the factors which are beyond the control of human hand. It is assumed that error components occur independently and at random such that

(*i*) $$E(X_e) = 0,$$

i.e., the error components increase as often as they decrease a measurement

(*ii*) $$\left.\begin{array}{l} E(X_\infty, X_e) = 0 \\ \text{and } E(X_{e_i} X_{e_j}) = 0, i \neq j \end{array}\right\}$$

i.e., error components are uncorrelated with the true values and the errors in other measurements.

Remark. These conditions may not always be true. In the absence of evidence to the contrary, we shall assume that they are satisfied. On taking expectation of both sides in (8·11), we get

$$E(X_t) = E(X_\infty)$$

which implies that the observed mean score equals the true mean score.

Next we segregate or split the total variance of a set of measures into two components, *viz.*, the true variance and the error variance. We have from (8·11),

$$\text{Var}\ (X_t) = \text{Var}\ (X_\infty + X_e) = \text{Var}\ (X_\infty) + \text{Var}\ (X_e),$$

covariance term vanishes because of (8·12*a*). This gives

$$s_t^2 = s_\infty^2 + s_e^2 \qquad \dots (8\cdot 13)$$

where s_t^2 = variance of the test score, s_∞^2 = True variance and s_e^2 = Error variance ... (8·13*a*)

Dividing both sides by s_t^2, we obtain $\quad 1 = \dfrac{s_\infty^2}{s_t^2} + \dfrac{s_e^2}{s_t^2}$

8·3·1. Definition of Reliability. *The reliability of any set of measurements is defined as that part of the variance which is true variance.*

If we write r_{tt}, for the coefficient of reliability of a test then, we have

$$r_{tt} = \frac{s_\infty^2}{s_t^2} \qquad \dots (8\cdot 14)$$

$$= 1 - \frac{s_e^2}{s_t^2} = 1 - \frac{\text{Error Variance}}{\text{Variance of raw scores}} \qquad \dots (8\cdot 14a)$$

Error Variance or Standard Error of Measurement. Solving equation (8·14*a*) for s_e, we get

$$s_t^2\, r_{tt} = s_t^2 - s_e^2 \quad \Rightarrow \quad s_e^2 = s_t^2\,(1 - r_{tt}) \quad \Rightarrow \quad s_e = s_t\,(1 - r_{tt})^{1/2} \qquad \dots (8\cdot 15)$$

This gives us the standard deviation of the error scores, also known as the *standard error of the measurement*.

Remarks **1.** Obviously reliability ranges from 0 to 1, *i.e.*,

$$0 \le t_{tt} \le 1 \qquad \dots (8\cdot 14b)$$

$r_{tt} = 1$, when $s_e = 0$. But since $E\ (X_e) = 0$, $s_e = 0$ iff $X_{e_i} = 0\ \forall\ i$, where X_{e_i} is the error score of the *i*th individual. Thus a test is perfectly reliable iff

$$X_{t_i} = X_\infty,\ \forall\, i\,,\ 2,\ \dots,\ n$$

and in this case raw scores are same as the true scores.

$r_{tt} = 0$, if $s_\infty = 0$, [from (8·14)] or equivalently from (8·14*a*) $s_e = s_t$

$$s_\infty = 0 \quad \Rightarrow \quad \text{Var}\ (X_\infty) = 0\ ,\ i.e.,\ \ X_\infty = \text{constant} = k\ (\text{say})$$

Thus $\quad r_{tt} = 0$ if $X_{t_i} = k + X_{e_i}$, $(i = 1, 2, \dots, n)$

where k denotes the true score for all i and in this case test is unreliable.

We also observe from (8·14*a*) that as the error variance increases, reliability decreases. The error variance to a certain extent can be controlled and reliability raised.

2. The important factors affecting the reliability of a test are :

(*i*) Length of the test. (*ii*) Range of talent. (*iii*) Ability level of the subject. (*iv*) Testing conditions.

8·3·2. Index of Reliability. Coefficient of reliability r_{tt} defined in (8·14) is merely an abstract idea. Operationally, it is some sort of self-correlation of a test.

Statistically, linear model (8·11) may be interpreted as the line of regression of X_t on X_∞, *i.e.*, we may regard the variates X_t and X_∞ representing a bivariate distribution in which X_t is a dependent variable and X_∞ is an independent variable. The Karl Pearson's correlation coefficient between X_t and X_∞, *i.e.*, between a series of obtained scores (X_t) and their corresponding theoretically true scores (X_∞) is known as the '*index of reliability*' and is written as $r_{t\infty}$. From regression theory we know that the standard error of estimate (as given by the line of regression of X_t on X_∞) is given by :

$$s_{t\infty} = s_t\,(1 - r_{t\infty}^{\;2})^{1/2} \qquad \ldots (8{\cdot}16)$$

But $s_{t\infty}^{\;2}$ is the same as the error variance $s_e^{\;2}$ given in (8·15). Hence on comparing (8·15) and (8·16), we get

$$1 - r_{tt} = 1 - r_{t\infty}^{\;2} \quad \Rightarrow \quad r^2_{t\infty} = r_{tt}, \quad i.e., \quad r_{t\infty} = \sqrt{r_{tt}} \qquad \ldots (8{\cdot}17)$$

This formula gives the index of reliability ($r_{t\infty}$) in terms of the coefficient of reliability (r_{tt}).

Remarks **1.** Since $0 \le r_{tt} \le 1$, from (8·17*a*), we conclude that $r_{t\infty}$ is numerically higher than r_{tt}, *i.e.*, $| r_{t\infty} | > | r_{tt} |$.

2. The highest correlation that can be obtained between a given test and any other measure is the one obtained between the test (raw) score (X_t) and their corresponding true scores X_∞. Very rarely, chance may lead to higher spurious correlation. Thus, *the statistic $r_{t\infty}$ is usually used to indicate the maximum correlation which a given test is capable of yielding in its present form.*

8·3·3. Parallel Tests. Two tests are said to be *parallel* if it makes no difference whether we used one or the other.

Let $X_{\infty ig}$ and $X_{\infty ih}$ be the true scores of the *i*th individual in two tests *g* and *h* respectively. If $X_{\infty ig} \ne X_{\infty ih}$, then it is unreasonable to assume that it makes no difference if we administer the test *g* or *h*. Thus, for tests *g* and *h* to be parallel, we assume that

$$X_{\infty ig} = X_{\infty ih}, \text{ for } \; i = 1, 2, \ldots, n \qquad \ldots (8{\cdot}18)$$

In other words, *for parallel tests, true scores of any individual should be same on either test.* Further we assume that

$$\text{Var}\,(X_{eg}) = \text{Var}\,(X_{eh}) \qquad \ldots (8{\cdot}19)$$

i.e., the error variances on the two tests should be same.

Equations (8·18) and (8·19) define parallel tests in terms of unknown quantities. By using equations (8·11), (8·12), (8·18) and (8·19), parallel tests can be defined in terms of the known quantities as follows :

we have from (8·11),

$$X_{tg} = X_{\infty g} + X_{eg}, \quad \text{and} \quad X_{th} = X_{\infty h} + X_{eh} \qquad \ldots (*)$$

On taking expectations and variances of both sides in (*) and using (8·12*a*) and (8·19) together with the fact that $X_{\infty g} = X_{\infty h}$ for parallel tests, we get for two parallel tests *g* and *h*.

$$E(X_{tg}) = E(X_{th}) \quad \text{and} \quad \text{Var}\,(X_{tg}) = \text{Var}\,(X_{th}) \qquad \ldots (8{\cdot}20)$$

which implies that the means and variances of the raw scores on two parallel tests are equal.

8·3·4. Methods of Determining Test Reliability. In this Section we shall briefly describe five methods in common use for estimating the reliability of tests.

(a) The Test-Retest Method. This method consists in submitting a group of individuals or candidates to a particular test and compiling their respective scores. After some time the same test is repeated on the same candidates and their scores are noted again. The two series of scores are arranged pairwise, a pair being the scores of the candidate in the two repititions of the test. Karl Pearson's coefficient of correlation between the two series is taken as the measure of coefficient of reliability.

Although simplest to administer and apply this method is subjected to several serious objections as given below :

(*i*) If the test is repeated immediately after the first, the scores are very likely to be improved on account of the memory effects, practice and confidence induced by familiarity with the test material.

(*ii*) On the other hand, if sufficient time is allowed to elapse before the second administration, the above effects, *viz*., memory, practice, etc., can be minimised or rather eliminated but in this case some other factors may creep in. For example, if the candidates are young children, their growth and maturity will, in general, increase the retest score if the test is given after a long time, *e.g.*, 6 months or more, thereby decreasing the reliability coefficient.

(*iii*) If a test contains novel features, it will all the more be influenced by practice as compared with a test containing routine and familiar items. Thus test-retest method is less reliable in the former case than in the latter case.

Owing to these practical difficulties, this method is not very popular.

(b) *Alternate or Parallel Forms Method.* Let there be a test '*A*' with n items 1, 2, ..., n. Manipulate a parallel test '*B*' (say) with n items 1′, 2 ′, ..., n'. The parallel forms method consists in administering two different but parallel forms of a test, say, A and B to a number of candidates and noting the corresponding two series of scores. The measure of reliability is provided by the coefficient of self-correlation between the two sets of scores as in the test-retest method.

Apparently, the parallel forms method appears to eliminate the draw-backs of test-retest method although some of them are inherent here also. This is a fairly reliable method of estimating test reliability in many practical situations provided the ability being measured does not change appreciably in the time interval that lapses between the administration of the two forms of the test. Hence this method is generally employed by the authors of the most standard psychological, educational and scholastic achievement tests. However, the method is not recommended if the trait being measured changes during the interval between two administration as, for example, in athletic skills test where practice is likely to improve the performance.

Another difficulty is the construction of the parallel test. The test items in the two forms should be such that they are neither too much identical (*w.r.t.* content, difficulty and form) nor too much unlike otherwise reliability will be favourably or unfavourably affected. The manipulation of parallel test items is a matter of controversy and hence this method is devoid of an international acceptance.

(c) *Split-half Method.* This method, as the name suggests, consists in breaking the original test into two equivalent halves and computing the correlation (r_{hh}) between the scores in half test. The coefficient of reliability r_{tt} for the whole test is calculated in terms of self-correlation r_{hh} of the half test by using Spearman-Brown formula.

$$r_{tt} = \frac{2r_{hh}}{1 + r_{hh}} \qquad \ldots (8{\cdot}21)$$

Merits. Split-half method is advisable when it is not possible to construct alternate forms of the test or repeat it, *e.g.*, in personality tests and in certain problems like picture drawing or puzzle solving. This method has been regarded by many as the best, its principal advantage being that all the experimental data for computing reliability of the test can be obtained in one occasion thus eliminating the memory and practice effects.

Demerits. The primary objection to this method is that there is no unique way of dividing the test into two halves and consequently there is no unique split half correlation and as such we cannot infer anything exact about reliability. This objection is basically true in case of speed tests (see remark below) or in tests where items are all of equal difficulty. However in case of power tests (see remark below) , the test items are arranged in increasing order of difficulty and thus splitting them into two halves with odd and even numbered items provides a unique estimate of reliability.

Remark. (*Speed and Power Tests*) Speed tests are those in which emphasis is laid on the speed or quickness with which the items can be answered by the candidates. In these tests, a time limit is imposed which is so short that all the items cannot be answered by all the candidates. Speed tests are of low difficulty level. On the other hand, in power tests, examinees are given sufficient time to try every item and the difficulty of items increases steadily.

Parallel forms or test-retest methods are to be used when speed is an important factor in the test score whereas in power tests split-half technique or the method of rational equivalence (discussed in the next section § 8·11·5) should be used.

(*d*) *The Rulon Method of Estimating Reliability.* Rulon gave the following formula for estimating the reliability from the scores on two halves of the same test :

$$r_{tt} = 1 - \frac{\sigma_d^2}{\sigma_t^2} \qquad \ldots(8{\cdot}22)$$

where σ_t^2 is the variance of the raw scores in the test and σ_d^2 is the variance of the difference of raw scores on the two halves of the test given by :

$$\sigma_d^2 = \frac{1}{n} \sum_{i=1}^{n} d_i^2 \qquad \ldots(8{\cdot}22a)$$

where d_i is the difference between the two scores of half tests for the *i*th individual.

Another formula which is much simple to apply is due to Guttman and is given by :

$$r_{tt} = 2\left[1 - \frac{\sigma_1^2 + \sigma_2^2}{\sigma_t^2}\right] \qquad \ldots(8{\cdot}23)$$

where σ_1^2 and σ_2^2 are the variances of the raw scores on the two halves.

If $\sigma_1^2 = \sigma_2^2$, *i.e.*, the two halves have equal raw score variances, then all the formulae in (8·21), (8·22), and (8·23) will give the same reliability coefficient otherwise the reliability coefficient given by (8·21) will be highest.

(*e*) *Method of Rational Equivalence or Kuder-Richardson Formula.* This method (due to Kuder-Richardson—1939) enables us to get an estimate of the coefficient of reliability free from the objections raised against the earlier methods. It stresses the intercorrelations of the items in the test and the correlations of the items with the test as a whole. Kuder-Richardson formula is based on the assumption that all the items of the test are of equal or nearly equal difficulty and intercorrelations.

The most accurate and useful of all, Kuder-Richardson formula for determining the test reliability in terms of the difficulty and intercorrelations of test items is :

$$r_{tt} = \frac{n}{n-1}\left[\frac{\sigma_t^2 - \sum_{i=1}^{n} p_i q_i}{\sigma_t^2}\right], \qquad \ldots(8{\cdot}24)$$

where r_{tt} is the reliability coefficient of the whole test,
n is the number of items in the test,
σ_t is the standard deviation of test scores,
p_i is the proportion of group answering *i*th test item correctly, and
$q_i = 1 - p_i$, $(= 1, 2, \ldots, n)$

If we are justified in assuming that all the items in the test have approximatety the same degree of difficulty then an approximation to (8·24) is provided by :

$$r_{tt} = \frac{n}{n-1}\left(\frac{\sigma_t^2 - n\,\bar{p}\,\bar{q}}{\sigma_t^2}\right); \qquad \ldots (8{\cdot}24a)$$

where $\bar{p}$ and $\bar{q}$ represent the average proportions of passing and failing examinees for each item respectively.

If $p_i = p \;\forall\; i = 1, 2, \ldots, n$, *i.e.*, if all the items are of equal difficulty and further approximation to (8·24) due to Forclich gives :

$$r_{tt} = \frac{n\sigma_t^2 - n^2pq}{(n-1)\sigma_t^2} \qquad \ldots (*)$$

But $n^2pq = np . n\,(1-p) = np\,(n-p) = M\,(n-M)$, because in case of binomial distribution np = Mean = M, (say). Hence substituting in (*), we get

$$r_{tt} = \frac{n\sigma_t^2 - M\,(n-M)}{(n-1)\,\sigma_t^2} \qquad \ldots (8{\cdot}24b)$$

This formula is used by teachers and others who want to determine quickly the reliability of short objective type class-room tests. It saves lot of time since it is based on the number of items in a test and the mean and standard deviation of test scores, and no correlations are to be calculated.

***Remarks* 1.** In formula (8·24*b*), we make the assumption that all the items are of equal difficulty, *i.e.*, the same proportion of persons (but not necessarily the same persons) pass each item. However, it has been seen in practice that formula (8·24*b*) provides a fairly good index of reliability even if the items vary considerably in difficulty.

2. It may be pointed out that the Kuder-Richardson formulae given above, which depend upon the single administration of a test, tend to underestimate the reliability of a test, the formula (8·24) most of all.

8·3·5. Effect of Test Length on the Reliability of the Test. Increasing the length of a test tends to increase its reliability. This increased reliability is determined by Spearman Brown prophecy formula :

$$r_{nn} = \frac{nr_{11}}{1+(n-1)r_{11}} \qquad \ldots (8{\cdot}25)$$

where r_{11} is the reliability of the given test of unit length,

r_{nn} is the correlation coefficient between n forms of the given test and n alternate forms (or the mean of n forms against the mean of n other forms), and

n is the number of times the length of a test is to be increased or decreased.

In particular, if we take n= 2, the reliability coefficient for doubled-test length becomes :

$$r_{22} = \frac{2r_{11}}{1+r_{11}} \qquad \ldots (8{\cdot}25a)$$

For the validity of formula (8·25), the items added in increasing the length of given test should be homogeneous, *i.e.*, (*i*) they should have about the same intercorrelation with the items in the test as those items have among themselves, and (*ii*) they should be of comparable difficulty with the items of the original test.

Again, the test should not be lengthened more than 6 or 7 times otherwise boredom, fatigue and loss of incentive, etc., may affect the results adversely.

***Remarks* 1.** We could use formula (8·25) if we had a long test, say, of 100 items that took 2 hours to administer. The idea is to cut this long test to 50 items so that it can be administered in half the time. The reliability for this test would be obtained by taking $n = \frac{1}{2}$ in (8·25). The problem could be reversed also. Suppose that we had a 25-item test of certain reliability $r_{11} = 0{\cdot}75$ (say). How long should it be lengthened to have a reliability coefficient $r_{nn} = 0{\cdot}90$ (say) ? For this using (8·26), as discussed below, we get

$$n = \frac{0{\cdot}90\,(1 - 0{\cdot}75)}{0{\cdot}75\,(1 - 0{\cdot}90)} = 3$$

Hence, the given test of reliability 0·75 will have to be 3 times as long, *i.e.*, should consist of 3 × 25 = 75 items to attain a reliability coefficient 0·90.

2. The prophecy formula (8·25) may also be used to determine how many times a test should be lengthened or repeated in other to obtain a test with specified reliability. For this solving (8·25) for n, we get

$$n = \frac{r_{nn}\,(1 - r_{11})}{r_{11}\,(1 - r_{nn})} \qquad \ldots (8{\cdot}26)$$

8·3·6. Effect of Different Ranges on the Reliability of the Test. The reliability coefficient of a test administered to two different groups, one of wide range and the other of relatively narrower spread cannot be measured directly. If the standard error of measurement s_e^2 remains constant irrespective of the range of ability in the group then from the formula :

$$r_{tt} = 1 - \frac{s_e^2}{s_t^2}\,,$$

we see that as the scatter (range) as measured by s_t^2 increases, r_{tt} increases and if s_t^2 decreases r_{tt} decreases. Thus the more heterogeneous the group is, the greater is the test variability and consequently, the reliability coefficient is higher.

If we know the reliability coefficient r_{00} of a test for a group with dispersion σ_0, then its reliability coefficient r_{nn} for another group with dispersion σ_n can be derived from the formula (on the assumption that the standard error of measurement is same for both the ranges)

$$\sigma_0\sqrt{1 - r_{00}} = \sigma_n\sqrt{1 - r_{nn}} \quad \Rightarrow \quad \frac{\sigma_n}{\sigma_0} = \left(\frac{1 - r_{00}}{1 - r_{nn}}\right)^{1/2} \qquad \ldots (8{\cdot}27)$$

Solving for r_{nn}, we obtain : $$r_{nn} = 1 - \frac{\sigma_0^2\,(1 - r_{00})}{\sigma_n^2}\,. \qquad \ldots (8{\cdot}27a)$$

***Example* 8·10.** *The reliability coefficient of a test of 50 items is 0·60. (a) How much the test should be lengthened to raise the self-correlation to 0·90?) (b) What effect will the (i) doubling, and (ii) tripling the test's length have upon the reliability coefficient ? (c) What is the reliability of a test having 125 comparable items ?*

Solution. (*a*) Here we are given $r_1 = 0{\cdot}60$ and $r_{nn} = 0{\cdot}90$. Substituting in (8·26), we get

$$n = \frac{0{\cdot}90\,(1 - 0{\cdot}60)}{0{\cdot}60\,(1 - 0{\cdot}90)} = 6$$

Hence the original test with $r_{11} = 0{\cdot}60$ should be 6 times as long to attain a reliability of 0·90, *i.e.*, it should contain 6 × 50 = 300 items.

(*b*) (*i*) Here $\qquad r_{11} = 0{\cdot}6,\ n = 2$ and $r_{nn} = ?$

Substituting in (8·25), $$r_{nn} = \frac{2 \times 0{\cdot}6}{1 + 0{\cdot}6} = 0{\cdot}75$$

(*ii*) When $n = 3$, $$r_{nn} = \frac{3 \times 0{\cdot}6}{1 + 2 \times 0{\cdot}6} = 0{\cdot}81$$

(*c*) Here $\qquad r_{11} = 0{\cdot}6$ and $n = \dfrac{125}{50} = 2{\cdot}5$, and $r_{nn} = \dfrac{2{\cdot}5 \times 0{\cdot}6}{1 + 1{\cdot}5 \times 0{\cdot}6} = 0{\cdot}79.$

Example 8·11. *A given test has a reliability coefficient of 0·8 and standard deviation of 20.*

(i) What is the maximum correlation which this test is capable of yielding as it stands ?

(ii) What is the S.E. of a score obtained on this test ?

(iii) What is the estimated reliability coefficient of this test in a group in which standard deviation is 15 ?

(iv) What proportion of the variance of the scores in this test is attributable to 'true' variance ?

Solution. (*i*) Maximum correlation is given by :

$$r_{0\infty} = \sqrt{r_{00}} = \sqrt{0{\cdot}80} = 0{\cdot}89$$

Hence test is capable of maximum correlation of 0·89.

(*ii*) Standard error of the score is given by :

$$\text{S.E.} = \sigma_0 \sqrt{1 - r_{00}} = 20 \times \sqrt{0{\cdot}20} = 20 \times 0{\cdot}4472 = 8{\cdot}944$$

(*iii*) In the usual notations, we are given $\sigma_0 = 20$, $r_{00} = 0{\cdot}80$, $\sigma_1 = 15$ and we are required to find r_{nn}. Using (8·27*a*), we get

$$r_{nn} = 1 - \frac{20^2\,(1 - 0{\cdot}80)}{15^2} = 1 - \frac{16}{45} = 0{\cdot}644$$

(*iv*) Since reliability is also defined as that part of the variance which is true variance, the proportion of the variance of the scores which is attributable to true variance is $r_{00} = 0{\cdot}80$, *i.e.*, 80%.

Example 8·12. *Show that the reliability ρ_k of a test at length k in terms of its reliability ρ_h at length h is given by :*

$$\rho_k = \frac{k\rho_h}{h + (k - h)\,\rho_h}$$

Solution. Let ρ_{11} be the reliability of the test of unit length ;

ρ_k : Reliability of the test of length k ;

ρ_h : Reliability of the test of length h.

Then by formula (8·25), *viz.*,

$$\rho_k = \frac{k\rho_{11}}{1 + (k-1)\,\rho_{11}} \quad \text{and} \quad \rho_h = \frac{h\rho_{11}}{1 + (h-1)\,\rho_{11}}$$

$$\Rightarrow \quad \frac{k}{\rho_k} = \frac{1 + (k-1)\,\rho_{11}}{\rho_{11}} = \frac{1}{\rho_{11}} + (k - 1) \quad \text{and} \quad \frac{h}{\rho_h} = \frac{1 + (h-1)\,\rho_{11}}{\rho_{11}} = \frac{1}{\rho_{11}} + (h - 1)$$

Subtracting (to eliminate ρ_{11}), we get

$$\frac{k}{\rho_k} - \frac{h}{\rho_h} = k - h \quad \Rightarrow \quad \frac{k}{\rho_k} = \frac{h + (k-h)\,\rho_h}{\rho_h} \qquad i.e., \quad \rho_k = \frac{k\,\rho_h}{h + (k - h)\,\rho_h}.$$