ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

The Realism of Reliability : The Contradiction between Definition and Estimation of Reliability

Surasak Amornrattanasak*

ABSTRACT

The purpose of this study is to show how the estimation of reliability from the real data is different from the definition of reliability. Theoretically, reliability is defined as the ratio of true score and observed score variances. From the definition of reliability, we can conclude that the reliability index ranged from 0 to 1. In fact, there are many cases in which the estimation of reliability from the real data is different from the conclusion. Subsequently, there is no reason to limit the reliability index between 0 to 1. Nowadays there is no reason to explain why the reliability index is less than -1.0. This is the main issue for all measurers to find the reasons why there is a significant contradiction between the definition and estimation of reliability.

Keywords : Reliability / Measurement Theory / Test Theory / Estimation of Reliability

Introduction

To measure and evaluate the academic achievement of learners, educational instruments are necessary. If all measuring instruments are of high validity and reliability, the results of educational measurement will be correct and will be correspond to the real educational situation. In contrast, if the educational instruments lack validity, it cannot be used to measure the abilities of learners correctly. For educational measurement and evaluation process, reliability is very important and necessary for all measuring instruments. The higher the reliability, the more stable the results of the measurement will be. In contrast, if the instruments lack reliability, the results become inconsistent and unreliable.

In addition, research instruments require high reliability. If the research instruments lack reliability, the research results will be incorrect and unreliable. Thus, finding the fact about reliability is very necessary. This is the reason why I proposed this article.

*Associate Professor, Faculty of Education, Ramkhamhaeng University

ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

The Definition of Reliability

Reliability can be defined as the degree of consistency between two measures of the same thing. This means that the scores from the two measures should be nearly the same or only slightly different. If the scores from each measurement are different, we cannot determine which one reports the accurate scores.

Ebel and Frisbie (1991) defined reliability as:

"The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalence test obtained independently from member of the same group."

Consequently, we can say that the classical definition of reliability is the idea of correlation and equivalence tests.

Theoretically, reliability is defined as the ratio of the true score and observed score variances (Mehrens and Lehmann, 1984). This can be expressed as follows:

$$r_{tt} = \frac{S_t^2}{S_x^2} \quad \text{..... (1)}$$

Since $x = t + e$

$$S_x^2 = S_t^2 + S_e^2$$

$$S_t^2 = S_x^2 - S_e^2$$

Thus

$$r_{tt} = 1 - \frac{S_e^2}{S_x^2} \quad \text{..... (2)}$$

Note: r_{tt} = reliability index

S_t^2 = true score variance

S_x^2 = observed score variance

S_e^2 = error score variance

We can summarize the reliability (r_{tt}) from equation (2) as follows:

1. $r_{tt} = 0$, where $S_e^2 = S_x^2$
2. $r_{tt} = 1$, where $S_e^2 = 0$
3. r_{tt} can never be negative because the variance of the error scores is always less than or equal to the variance of the observed scores ($S_e^2 \leq S_x^2$, since $S_x^2 = S_t^2 + S_e^2$)

From the above definition of reliability, Mehrens and Lehmann (1984); Allen and Yen (2002) concluded that the reliability index must be positive because each variance value is always positive and ranged from 0 to 1.

ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

This is empirically confirmed by Malhotra (2004) that Cronbach's alpha is a reliability coefficient that measures inter-item reliability or the degree of internal consistency / homogeneity between variables measuring one construct / concept i.e. the degree to which different items measuring the same variable attain consistent results. This coefficient varies from 0 to 1 and a value of 0.6 or less generally indicates unsatisfactory internal consistency reliability.

The Contradiction between Definition and Estimation of Reliability

In fact, there are many cases in which the estimation of reliability from the real data is different from the above conclusion.

Since $S_x^2 = S_t^2 + S_e^2$ depends on the assumption that there is no correlation between true scores and error scores ($r_{te} = 0$), the variance of the error scores must be less than or equal to the observed. This causes the reliability to violate the assumption.

Conversely, if the correlation coefficient between true scores and error scores is not zero, the equation $S_x^2 = S_t^2 + S_e^2$ is not absolutely true. The variance of observed scores can be represented as follows:

$$S_x^2 = S_t^2 + S_e^2 + 2\, r_{te}\, S_t\, S_e$$

Consequently, if the correlation coefficient between true scores and error scores is negative, it's possible that $S_e^2 > S_x^2$. In this case, the reliability index in equation (2) will be negative.

Another reason why the reliability violates this assumption is due to the definition from equation (1), which is derived from the correlation between observed scores on two parallel tests. Magnusson (1967) showed how to derive the estimation of the reliability index as follows:

$$r_{x_g x_h} = \frac{\sum x_g x_h}{N S_{x_g} S_{x_h}} \qquad \text{..... (3)}$$

where x_g is observed score from test G

x_h is observed score from test H

Finally, the result would be

$$r_{x_g x_h} = \frac{S_t^2}{S_x^2} \qquad \text{..... (4)}$$

Equation (4) has become the definition of reliability nowadays. However, some measurers conclude that the reliability index varies from 0 to 1. This conclusion is not absolutely true, because the definition of reliability from equation (4) is derived from the correlation between observed scores on two parallel tests. In brief, the correlation coefficient would vary from -1.0 to +1.0 so the reliability index can be negative.

ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

Magnusson (1967) derived Cronbach's coefficient alpha, the reliability estimated by measure of internal consistency, as follows:

$$S_x^2 = \sum S_i^2 + 2\sum r_{ik} S_i S_k \quad , \quad i > k \qquad \text{..... (5)}$$

The first term of equation (5) is variance with n terms while the last term is covariance with n (n-1) terms.

If each item is parallel, the probability of answering correctly should be equal. In addition, if the inter-correlation among each item is equal, S_i S_k by S_i^2 can be substituted.

Thus, the summation of the last term in equation (5) can substitute by n (n-1) $\bar{r}_{ik}$ $\bar{S}_i^2$

Equation (5) can be rewritten as follows:

$$S_x^2 = \sum S_i^2 + n(n-1)\bar{r}_{ik}\bar{S}_i^2 \qquad \text{..... (6)}$$

Since $n\bar{S}_i^2 = \sum S_i^2$ and can be substituted into equation (6), we get

$$\bar{r}_{ik} = \frac{S_x^2 - \sum S_i^2}{(n-1)\sum S_i^2} \qquad \text{..... (7)}$$

Equation (7) $\bar{r}_{ik}$ is the correlation coefficient between item i and item k, which is the reliability index of one item. If we want to obtain an estimate of reliability based on a full-test which has n items, it is necessary to correct the correlation of one item to the full-length correlation. This is done with the help of the Spearman-Brown formula as follows:

$$\begin{aligned} r_{tt} &= \frac{n\bar{r}_{ik}}{1+(n-1)\bar{r}_{ik}} \\ &= \frac{n\left(S_x^2 - \sum S_i^2\right)}{(n-1)\sum S_i^2} \frac{1}{1+(n-1)\left[\left(S_x^2 - \sum S_i^2\right)/(n-1)\sum S_i^2\right]} \\ &= \frac{n}{(n-1)} \frac{S_x^2 - \sum S_i^2}{\sum S_i^2} \frac{1}{1+\left(S_x^2 / \sum S_i^2\right)-1} \\ &= \frac{n}{(n-1)} \frac{S_x^2 - \sum S_i^2}{\sum S_i^2} \frac{\sum S_i^2}{S_x^2} \\ &= \frac{n}{(n-1)} \frac{S_x^2 - \sum S_i^2}{S_x^2} \\ &= \frac{n}{(n-1)} \left[1 - \frac{\sum S_i^2}{S_x^2}\right] \qquad \text{..... (8)} \end{aligned}$$

ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

Equation (8) is a general formula to estimate the reliability index, which we call Cronbach's coefficient alpha.

Cronbach's coefficient alpha is derived from the variance of observed scores in equation (5).

Consequently, if the term of covariance is negative, $\sum S_i^2$ can be greater than S_x^2. This causes the reliability index to be negative.

Furthermore, the equation (8) is derived from the correlation between observed scores on two parallel items and corrects with the help of Spearman-Brown formula. So there is no reason to conclude that the reliability index is ranged from 0 to 1 as the correlation coefficient would vary from -1.0 to +1.0 as previously mentioned.

Mehrens and Lehmann (1984) proposed that three methods of estimating reliability are applied using the Pearson Product Moment: the measure of stability (Test-retest method), the measure of equivalence (Parallel test method), and the measure of internal consistency (Split-half method). In this manner, we can find the correlation coefficient between two sets of observed scores.

Subsequently, there is no reason to limit the reliability index of these three methods between 0 to 1 because the correlation coefficient can vary from -1.0 to +1.0.

In addition, KR-20 and KR-21, developed by Kuder and Richardson, are two of the most widely accepted methods for estimating reliability.

KR-20 and KR-21 are actually a special case of Cronbach's coefficient alpha. The KR-20 formula is applicable only with tests scored dichotomously (0 or 1). 1 is for a correct answer and 0 is for an incorrect answer.

If the items are scored dichotomously (0, 1), we can prove that $S_i^2 = pq$

Thus $$r_{tt} = \frac{n}{(n-1)}\left[1 - \frac{\sum pq}{S_x^2}\right] \qquad \text{.....(9)}$$

Equation (9) is the KR-20 formula. If the difficulty of each item (p) is equal, we get $\sum pq = n\overline{pq}$. Substituting into (9), we get

$$r_{tt} = \frac{n}{(n-1)}\left[1 - \frac{n\overline{pq}}{S_x^2}\right] \qquad \text{.....(10)}$$

$$= \frac{n}{n-1}\left(1 - \frac{\frac{n\sum p}{n} \cdot \frac{\sum(1-p)}{n}}{S_x^2}\right)$$

ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

$$= \frac{n}{n-1}\left(1 - \frac{\sum p \cdot \frac{\sum 1 - \sum p}{n}}{S_x^2}\right)$$

$$= \frac{n}{n-1}\left(1 - \frac{\frac{\overline{X}(n-\overline{X})}{n}}{S_x^2}\right)$$

$$= \frac{n}{n-1}\left(1 - \frac{\overline{X}(n-\overline{X})}{nS_x^2}\right) \qquad \text{.....(11)}$$

Equation 10 and 11 are KR-21 formula, and it is less than the estimate of KR-20. This means that the reliability index of KR-21 is less than or equal to KR-20.

The reason why the reliability indices of KR-20 and KR-21 are negative can be explained by the same reasoning as Cronbach's coefficient alpha.

Similarly, Hoyt estimated the reliability by applying analysis of variance, which can be expressed as given below:

$$r_{tt} = 1 - \frac{MS_{residual}}{MS_{persons}} \qquad \text{.....(12)}$$

Hoyt's reliability index from equation (12) conforms to the definition of reliability ($r_{tt} = 1 - \frac{S_e^2}{S_x^2}$).

Therefore, Hoyt's reliability index can be negative because there is no condition that $MS_{persons}$ must be greater than $MS_{residual}$. If $MS_{residual}$ is greater than $MS_{persons}$, the reliability index will be negative.

Subsequently, Hoyt's method to estimate reliability should not be ranged between 0 to 1.

If the reliability index of an instrument is negative, this means that the instrument cannot be used to measure any trait.

However, there is no reason to explain why the reliability index using KR-20 or Cronbach's coefficient alpha is less than -1.0. It can be only summarized that the formulas above, KR-20 and Cronbach's coefficient alpha, are derived from the correlation between two sets of scores and leads to a reliability index of not less than -1.0.

ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

Results from Real Data

As in examples 1 and 2, it will support the idea that the reliability index from real data can be negative and less than -1.0

Example 1

item / person	1	2	3	4	5	X	X^2
1	1	1	0	1	1	4	16
2	1	1	0	1	0	3	9
3	0	1	1	1	1	4	16
4	1	0	1	1	1	4	16
5	1	0	1	0	1	3	9
6	0	1	1	0	1	3	9
7	0	1	1	1	0	3	9
8	0	0	1	1	1	3	9
9	1	0	1	0	0	2	4
10	1	0	0	1	0	2	4
p	.6	.5	.7	.7	.6		
q	.4	.5	.3	.3	.4		
pq	.24	.25	.21	.21	.24		

Σpq = 1.15

ΣX = 31 N = 10 $\overline{X}$ = 3.1 n = 5

ΣX^2 = 101

S_x^2 = .49

KR20 : $r_{tt} = \frac{n}{(n-1)}\left(1-\frac{\Sigma pq}{S_x^2}\right)$

$= \frac{5}{4}\left(1-\frac{1.15}{.49}\right)$

= 1.25 (-1.34)

= -1.675

ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

Example 2

item / person	1	2	3	4	5	X	X^2
1	5	1	3	4	4	17	289
2	5	1	3	4	3	16	256
3	4	2	3	3	3	15	225
4	4	2	3	3	4	16	256
5	3	3	3	5	2	16	256
6	3	3	4	5	5	20	400
7	2	4	4	5	5	20	400
8	2	4	4	5	2	17	289
9	1	5	5	4	3	18	324
10	1	5	4	4	4	18	324
$\overline{X}$	3.0	3.0	3.6	4.2	3.5	17.3	
S^2	2.222	2.222	.489	.622	1.167	2.900	

$$\alpha = \frac{n}{(n-1)}\left(1-\frac{\Sigma S_i^2}{S_x^2}\right)$$

$$= \frac{5}{4}\left(1-\frac{6.722}{2.90}\right)$$

$$= 1.25 \quad (1 - 2.138)$$

$$= 1.25 \ (-1.138)$$

$$= -1.423$$

Recommendation

The reliability indices are derived from the correlation between two sets of scores, so the correlation coefficient (reliability index) would vary from -1.0 to +1.0. If the reliability index is derived from the theoretical definition, the reliability index should range from 0 to 1 because reliability is the ratio of true score and observed score variance. This definition is under some limited conditions. When it is not, the assumptions of a valid estimation procedure have been violated. This causes the values of reliability are outside of the bound as mentioned above.

Any value less than zero should be a red flag and should not be interpreted. To get large negative values from Cronbach's coefficient alpha, it means there is strong negative covariances and very little variances. This means we have items with negative item- total correlations. Such items are either mis-keyed or they violate the assumption of

ปีที่ 21 ฉบับที่ 3 กันยายน - ธันวาคม 2558 ISSN 2408 - 0845

unidimensionality. Mostly in real testing situations this indicates a problem with the items or a more unusual situation in which the assumptions of Cronbach's coefficient alpha are violated.

From example 1 and example 2, the reliability indices are less than -1.0. In this case, there is no reason to explain why the reliability is less than – 1.0. Any explanations of this phenomenon would require further study. This comes to the main issue for all measurers to find a reason why there is a significant contradiction between the definition and estimation of reliability.

If we can't find any reasons to explain the above phenomenon, this means that we find the new body of knowledge. Consequently, the concept of teaching educational measurement and research must be changed, especially the reliability index should not limit from 0 to 1 as some measurers have mentioned above.

References

Allen, M.T., & Yen, W.N. (2002). **Introduction to measurement theory**. Long Grove, Illinois : Waveland Press.

Ebel, R. L., & Frisbie, D. A. (1991). **Essentials of educational measurement**. (5 th ed.). New Jersey : Prentice Hall.

Magnusson, D. (1967). **Test theory**. London : Addison-Wesley.

Malhotra, N.K. (2004). **Marketing research : an opplied orientation**. (4 th ed). New Jersey : Pearson Education.

Mehrens, W. A., & Lehmann, I. J. (1984). **Measurement and evaluation in education and psychology**. (3 rd ed). New York : CBS College Publishing.