Original Research Paper Education

Item-Total Correlation as the Cause for the Underestimation of the Alpha Estimate for the Reliability of the Scale

Jari Metsämuuronen University of Helsinki, Finland Finnish Education Evaluation Centre, Finlan

ABSTRACT *By modifying the basic formulae of the alpha coefficient for estimating the scale reliability, a simple coefficient is derived where the item-total correlation is seen. While knowing that the alpha coefficient gives the lower bound for the reliability except the case of (essential) tau-equivalency, the reason for the underestimation is in the process of calculating the item-total correlation. The article shows why and how much the item-total correlation underestimates the item discrimination in the deterministically discriminating dataset; even in the optimally constructed dataset, the underestimation may be 13%.*

KEYWORDS : Reliability; Alpha coefficient; Item discrimination; Item-total correlation; Discrimination Index

1. Introduction

During the recent years, the discussions about different aspects of the test score or scale reliability have been active (e.g. Graham, 2006; Raykov & Marcoulides, 2016; 2015; 2013; 2012; Raykov & Traynor, 2016; Raykov, West, & Traynor, 2015). Specifically, the researchers have been productive around the concept of *maximal reliability* within the SEM analysis (see Tenko Raykov's and his colleagues works in https://msu.edu/~raykov/Raykov_short_vitae.pdf). The findings related with the reliability of factor scores have been done independently also within the traditional exploratory factor analysis (EFA) (e.g. Tarkkonen, 1987; Vehkalahti, 1995; 2000) but the connection to SEM analysis have gained greater popularity.

From the 1937 on (Kuder & Richardson, 1937; Gulliksen, 1950; Cronbach, 1951), the classical alpha estimator (α) for the reliability have gained popularity. According to Hogan, Benjamin, and Brezinski (2000), Graham (2006), and Yang and Green (2011), the alpha estimate is the most used estimate for reliability for the unweighted scores. It is known that alpha is equal to reliability in conditions of (essential) tau-equivalence, that is, unless the true scores (taus) in the scale components are (essentially) equivalent, the alpha estimate for the scale reliability underestimates the composite reliability coefficient (see Guttman, 1945; Gulliksen, 1950; Kristoff, 1974; Novick & Lewis, 1967; Lord & Novick, 1968, 87–90; ten Berge & Zegers, 1978; Raykov, 1997; Vehkalahti, 2000; Raykov, Dimitrov, & Asparouhov, 2010; Metsämuuronen, 2017; 176–177) and the underestimation may be substantial (Raykov, 1997). This has led to the discussion about the *greatest* lower bound to reliability (e.g. Jackson & Agunwamba, 1977; Callender & Osburn, 1979; ten Berge, Snijders, & Zegers, 1981; ten Berge & Sočan, 2004) as well as maximal reliability. Graham (2006) and Raykov (1997) showed that the larger the violation of tau-equivalence in the test, the more alpha coefficient underestimates score reliability. Raykov (1997) showed also that the underestimation is less vulnerable with the test with a greater number of items.

This article elaborates the underestimation characteristics of the alpha estimator. A specific focus is the connection of alpha coefficient to the item discrimination, and more specifically, to the item-total correlation. The algebraic connection of the item-total correlation and alpha estimator is known from Lord and Novick (1968, 331, see formula 1). This leads to an obvious conclusion connected to the underestimation in test score reliability: the reasons and magnitude for the underestimation of the reliability by using the alpha estimate have to do something with the item-total correlation. Though the treaty is not restricted to the dichotomous case, focusing on the point-biserial correlation in the dichotomous dataset makes the discussion easier to adopt. This article shows that the point-biserial correlation always underestimates the item discrimination even in the deterministically discriminating, Guttman-type of dataset. Only when all the items of the dichotomous dataset are equally difficult, that is, in the (essentially) tau-equivalent case, $\rho_{gX}=1$. In all other cases, the product-moment correlation coefficient underestimates the item discrimination and, consequently, the alpha estimate underestimates the reliability even though the dataset would discriminate the test scores and cases from each other perfectly.

2. Item discrimination in the classical estimators of reliability

Recall the Lord and Novick (1968, p. 331) formula for alpha reliability:

$$\alpha=\frac{k}{k-1}\left(1-\frac{\sum_{g=1}^{k}\sigma_g^2}{\sigma_X^2}\right)=\frac{k}{k-1}\left(1-\frac{\sum_{g=1}^{k}\sigma_g^2}{\left(\sum_{g=1}^{k}\sigma_g\rho_{gX}\right)^2}\right) \quad (1)$$

where *k* refers to the number of items, σ_g^2 refers to the variances of single items *g*, σ_X^2 refers to the variance in the score, and ρ_{gX} refers to the item-total correlation between item *g* and the score *X*. The coefficient is algebraically identical with the original formula (Gulliksen, 1950), but the item discrimination ρ_{gX} is seen in the formula. The formula is not restricted to dichotomous cases. Recall also Ebel's (1967) formula, based on Stanley (1964):

$$\rho_{EBEL} = \frac{k}{k-1}\left(1 - \frac{6\sum_{g=1}^{k}\sigma_g^2}{\left(\sum_{g=1}^{k} DI\right)^2}\right) \quad (2)$$

where *DI* refers to Kelley's (1939) Discrimination Index, another estimator for the item discrimination. Ebel's coefficient is not in the general use though.

While knowing from (1) that ρ_{gX} is the element needed for estimating the test score reliability and when knowing that the alpha estimate always underestimates the reliability, the reason for the underestimation has to do something with the ρ_{gX} . It will be seen that, even with the deterministically discriminating Guttman type of items, ρ_{gX} always underestimates the item discrimination except the (essentially) tau-equivalent situation of strictly equal item difficulties in the dichotomous dataset. Though the treaty in what follows is not restricted to the achievement testing with 0/1 items, such wordings of "test-takers", "wrong answer", and "correct answer" are used to keep the discussion more practical.

3. Underestimation of reliability caused by the underestimation in point-biserial correlation in the dichotomous dataset

In the dichotomous dataset, the point-biserial correlation can be expressed as follows:

$$\rho_{gX} = (M^+ - M^-)\frac{\sigma_g}{\sigma_X} \quad (3)$$

where M^+ refers to the average score of the test-takers giving the correct answer (scoring 1) – let's call this group as *upper group* – and M^- refers to the average score in the group giving the incorrect answer (scoring 0) let's call this group the *lower group* – and σ_g refers to the standard deviation of the item *g,* and σ_X refers to the standard deviation of the score *X*.

Let us denote the number of the cases in the upper and lower groups by (non-symmetrically) N^+ and N^-. Obviously, the total number of all cases (*N*) is the sum of both groups:

$$N = N^+ + N^- \quad (4)$$

Because all the 1s are in the upper group

$$\frac{N^+}{N} = p \quad (5)$$

and, because of (4) and (5),

$$\frac{N^-}{N} = 1 - p \tag{6}$$

By denoting is the grand mean in the total score with *GM*, the variance of the test score can be manipulated as follows:

$$\begin{aligned}
\sigma_X^2 &= \frac{1}{N}\left(\sum_{i=1}^{N^+}(X_i^+ - GM)^2 + \sum_{i=1}^{N^-}(X_i^- - GM)^2\right) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N^+}(X_i^+ - M^+)^2 + N^+(M^+ - GM)^2 + \sum_{i=1}^{N^-}(X_i^- - M^-)^2 + N^-(M^- - GM)^2\right) \\
&= \frac{1}{N}\left(N^+\sum_{i=1}^{N^+}\frac{(X_i^+ - M^+)^2}{N^+} + N^-\sum_{i=1}^{N^-}\frac{(X_i^- - M^-)^2}{N^-} + N^+(M^+ - GM)^2 + N^-(M^- - GM)^2\right) \\
&= \frac{1}{N}\left(N^+\sigma_X^{2+} + N^-\sigma_X^{2-} + N^+(M^+ - GM)^2 + N^-(M^- - GM)^2\right) \\
&= \left(p\sigma_X^{2+} + (1-p)\sigma_X^{2-} + p(M^+ - GM)^2 + (1-p)(M^- - GM)^2\right) \text{ because of (5) and (6)}
\end{aligned} \tag{7}$$

.

The term $(M^+ - M^-)\sigma_g$ is manipulated as follows

$$\left(M^+ - M^-\right)\sigma_g = \sigma_g\left[\left(M^+ - GM\right) - \left(M^- - GM\right)\right] \tag{8}$$

Hence, the item total correlation can be expressed in the form

$$\rho_{gX} = \frac{\sigma_g\left[\left(M^+ - GM\right) - \left(M^- - GM\right)\right]}{\sqrt{p\sigma_X^{2+} + (1-p)\sigma_X^{2-} + p(M^+ - GM)^2 + (1-p)(M^- - GM)^2}} \tag{9}$$

To show some practical notes of the underestimation of point-biserial correlation in relation with the underestimation of reliability, two theoretical, deterministically discriminating datasets are used (see Tables 1 and 2). Both datasets consists of 11 items and 12 cases. Because of the ultimate discrimination, the estimates of the reliability should give the value 1. However, this will not happen if the items are not (essentially) tau-equivalent as discussed above.

3.1 Coefficient alpha in the (essentially) tau-equivalent situation

The point-biserial correlation (9) can reach the value $\rho_{gX} = 1$ only with one condition: when the variances of the score in both upper and lower group are equally $\sigma_X^{2+} = \sigma_X^{2-} = 0$ as in Table 1.

Table 1. Deterministically discriminating dataset with equal item difficulties

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	score
	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	1	1	1	1	1	1	1	1	1	11
	1	1	1	1	1	1	1	1	1	1	1	11
	1	1	1	1	1	1	1	1	1	1	1	11
	1	1	1	1	1	1	1	1	1	1	1	11
p	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33	
1-p	0,67	0,67	0,67	0,67	0,67	0,67	0,67	0,67	0,67	0,67	0,67	
p(1-p)	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22	
σ_{gX} = sqrt(p(1-p))	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,47	
ρ_{gx}	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	
σ_{gX} x ρ_{gx}	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,47	0,47	
DI33%	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	

The equal (observed) item difficulties do not necessarily mean that the true scores would be equal. Let us assume that the true scores *are* equal in Table 1. The assumption of tau-equivalency in relation with the alpha estimator comes from the original derivation of the alpha coefficient (Kuder & Richardson, 1937) based on generalizing the idea of parallel tests from the classical Spearman-Brown prophecy formula (Spearman, 1910; Brown, 1910) to the situation that all the individual items are parallel "tests". The parallelism implies that the true scores of the parallel tests correlate perfectly: $\rho_{\tau_g,\tau_h} = 1$. It also means that the *observed* scores of the parallel tests correlate perfectly $\rho_{gh} = 1$ (Gulliksen, 1950, 13–14; see also Metsämuuronen, 2017, proof 1). In the dichotomous dataset, the latter leads to the assumption of the equal item difficulties as well as the equal variances of the items.

The dataset in Table 1 shows (one of the possible) tau-equivalent case just as an example. The dataset fulfills the technical requirement of parallelism of the items: $\rho_{gh} = 1$, $\pi_g = \pi_h$, and $\sigma_g^2 = \sigma_h^2$. In this data structure, also the item total correlations are identically $\rho_{gX} = 1$. The dataset is formed so that also the *DI* = 1.

In practical terms, the pattern of equal difficulty levels means that, in both the upper and lower group, there is only one value in the score: 0 in the lower group and *k* in the upper group. Then, because of ultimate symmetricity in the scores,

$$p\left(M^+ - GM\right) = -(1-p)\left(M^- - GM\right) \tag{10}$$

and

$$\left(p\left(M^+ - GM\right)\right)^2 = p^2\left(M^+ - GM\right)^2 = (1-p)^2\left(M^- - GM\right)^2 \tag{11}$$

Because of (9), (10) and (11), the item-total correlation is

$$\rho_{gX} = \frac{\sigma_g \cdot \left(\frac{1}{1-p}\right)\left(M^+ - GM\right)}{\sqrt{\frac{p(1-p)}{(1-p)^2}\left(M^+ - GM\right)^2}} = \frac{\sigma_g \cdot \left(\frac{1}{1-p}\right)\left(M^+ - GM\right)}{\left(\frac{\sigma_g}{1-p}\right)\sqrt{\left(M^+ - GM\right)^2}} = 1 \tag{12}$$

In the case of essential tau-equivalence, the deterministically discriminating dataset produces the value α = 1 by the alpha coefficient (1) because $\rho_{gX} = 1$ and equal item variances. The Ebel formula (2) gives the value of 0.97, which is obviously an underestimation taking into account that the item discrimination measured by *DI* equals 1. In the practical settings, then, the perfect value for the alpha estimate would refer to the specific situation where the test takers can be deterministically divided in the two groups of zeros and maximum score. However, this data structure is an ultimately theoretical one – it is a very demanding task to find a large number of parallel items for a test (Raykov, Dimitrov, & Asparouhov, 2010). Regardless the theoretical essence of the data in Table 1, the structure of the dataset may be seen as the latent structure for the assessment tests based on standards. When our intention is to construct a test to assess a certain level of proficiency we may expect the test items to be relevant for this level. Hence, we may be willing to select test items which are more or less tau-equivalent.

3.2 Coefficient alpha in the non-tau-equivalent situation

The other theoretical dataset (Table 2) is also a deterministically discriminating one. However, it is a Guttman type of dataset (Guttman, 1950, Linacre & Wright, 1996), that is, items are patterned with a string of 0s followed by a string of 1s when the respondents are ranked in an ascending order by the score and the dataset forms an incremental structure of the difficulty levels.

Table 2. Deterministically discriminating dataset with non-equal variances

	v1	v2	v3	v4	v5	**v6**	v7	v8	v9	v10	v11	sum
	0	0	0	0	0	**0**	0	0	0	0	0	0
	0	0	0	0	0	**0**	0	0	0	0	1	1
	0	0	0	0	0	**0**	0	0	0	1	1	2
	0	0	0	0	0	**0**	0	0	1	1	1	3
	0	0	0	0	0	**0**	0	1	1	1	1	4
	0	0	0	0	0	**0**	1	1	1	1	1	5
	0	0	0	0	0	**1**	1	1	1	1	1	6
	0	0	0	0	1	**1**	1	1	1	1	1	7
	0	0	0	1	1	**1**	1	1	1	1	1	8
	0	0	1	1	1	**1**	1	1	1	1	1	9
	0	1	1	1	1	**1**	1	1	1	1	1	10
	1	1	1	1	1	**1**	1	1	1	1	1	11
p	0,08	0,17	0,25	0,33	0,42	**0,50**	0,58	0,67	0,75	0,83	0,92	
1-p	0,92	0,83	0,75	0,67	0,58	**0,50**	0,42	0,33	0,25	0,17	0,08	
p(1-p)	0,08	0,14	0,19	0,22	0,24	**0,25**	0,24	0,22	0,19	0,14	0,08	
σ_{gX} = sqrt(p(1-p))	0,28	0,37	0,43	0,47	0,49	**0,50**	0,49	0,47	0,43	0,37	0,28	
ρ_{gx}	0,48	0,65	0,75	0,82	0,86	**0,87**	0,86	0,82	0,75	0,65	0,48	
σ_{gX} x ρ_{gX}	0,13	0,24	0,33	0,39	0,42	**0,43**	0,42	0,39	0,33	0,24	0,13	
DI33%	0,33	0,67	1,00	1,00	1,00	**1,00**	1,00	1,00	1,00	0,67	0,33	
varM+	10,00	8,25	6,67	5,25	4,00	**2,92**	2,00	1,25	0,67	0,25	0,00	
varM-	0,00	0,25	0,67	1,25	2,00	**2,92**	4,00	5,25	6,67	8,25	10,00	
σ^2_{gX} (1-mean$_{pgh}$)	0,04	0,07	0,10	0,11	0,13	**0,13**	0,13	0,11	0,10	0,07	0,04	

Though the data structure in Table 2 produces, as Table 1, a score which ultimately discriminates the cases from each other, the structure leads to an underestimation of the reliability though the values are high because of an obviously unequal item difficulties, item variances and consequently, the unequal inter-item correlations. The standard formula of KR20 in (1) produces the value of 0.92.

Let us think about the highlighted item in the middle (*v6*). This is an item with $p = 0.50$ and the perfect discrimination in a sense that the item can deterministically discriminate the cases who gave the correct answer from those who gave the incorrect one while also the score can discriminate all the cases from each other. In this item, the point-biserial correlation is the highest ($\rho_{gX} = 0.87$). Why the value is not perfectly 1? Obviously, because of the mathematical procedure in calculating the correlation.

Let us elaborate the item *v6*. The symmetrically incremental structure of the data determines three things. First, the number of cases in both upper (+) and lower group (-) are the same ($N^+ = N^- = N/2$). Second, the variances of the score in both parts are identical:

$$\sigma_X^{2+} = \sigma_X^{2-} \tag{13}$$

Third, since there are only two means and the number of cases in the halves is the same, the means of the scores in the lower half (M^-), the upper half (M^+) and the grand mean (GM) of the reduced data are connected. Namely, because of (11) and knowing that $p = 0.50$:

$$\left(M^- - GM\right) = -\left(M^+ - GM\right) \tag{14}$$

and, because of (11)

$$\left(M^- - GM\right)^2 = \left(M^+ - GM\right)^2. \tag{15}$$

Because $p = 0.50$ in the item and because of (9), (14) and (15), the item-total correlation is:

$$\begin{aligned}\rho_{gX} &= \frac{\sigma_g\left[\left(M^+ - GM\right) - \left(M^- - GM\right)\right]}{\sqrt{p\sigma_X^{2+} + (1-p)\sigma_X^{2-} + p(M^+ - GM)^2 + (1-p)(M^- - GM)^2}} \\ &= \frac{0.5 \cdot 2\left(M^+ - GM\right)}{\sqrt{\sigma_X^{2+} + 0.5(M^+ - GM)^2 + 0.5(M^- - GM)^2}} \\ &= \frac{\left(M^+ - GM\right)}{\sqrt{\sigma_X^{2+} + (M^+ - GM)^2}} = \frac{\left(M^- - GM\right)}{\sqrt{\sigma_X^{2-} + (M^- - GM)^2}}\end{aligned} \tag{16}$$

Formula (16) shows, in a simple way, the same as noted above: the value of ρ_{gX} cannot reach the perfect 1 unless the variance in the sub-groups + or – is equally $\sigma_X^{2+} = \sigma_X^{2-} = 0$. Formula (16) also means that the higher is the variance in the upper or lower group, the more drastic is the underestimation of item discrimination even in the deterministically discriminating dataset. In the case, the underestimation is (1 – 0.87 = 0.13), that is, 13 percent – the value would be the same with bigger datasets also if the structure stayed the same.

Though the dataset in Table 2 is a theoretical one, it is a kind of ultimate structure seen behind many achievement tests with the incremental difficulty level; usually the test items in the static achievement testing are selected so that the test starts with easy items and ends with the demanding ones. As seen, in these kinds of situations, the underestimation of the reliability may be substantial as noted by Raykov (1997); at the best, it is 13% of the maximum. The reason for the underestimation of the reliability lies in the underestimation in item discrimination – caused by the mathematical procedure in calculating the correlation coefficient.

4. Discussion

The article shows that the reason for the underestimation in reliability is bound to underestimation in item-total correlation. Point-biserial correlation underestimates the item discrimination except in the (essentially) tau-equivalent situation. Even in the deterministically discriminating dataset, the correlation cannot reach the value 1 if there was variance in the sub-scores of those who reached 1

and those who reached 0. Point-biserial correlation is also very unstable indicator of item discrimination which also affect the underestimation in reliability.

Usually we tend to think that the error in testing comes from the random error and systematic error. The results revealed a kind of source of a systematic error in alpha estimator caused by the computing procedure of correlation coefficient. In many practical testing settings the alpha estimate may be valid as giving the lower bound of the reliability. However, because of knowing the reason for and amount of the underestimation, it may be possible to create a correction factor which takes into account the technical challenges in point-biserial correlation. Would it be possible utilize this knowledge in the latent trait modeling? Maybe it could be possible to correct the estimate of the item-total correlation? It may be worth noting that Metsämuuronen (2017, proof 7) showed that the item-*rest*-correlation (Henrysson, 1963) underestimates the item-discrimination even more than the item-total correlation in the specific situation of deterministically discriminating dataset with the incremental structure. Hence, this is not the way to go.

Another option is to start to think about the item discrimination from a different perspective: Why the item discrimination should be based on the actual values of the test score in the first place? Would it be possible to use only the order of the cases as the *DI* is using – but using the whole dataset in the calculations? This path may lead us to another kind of estimators for the item discrimination and reliability.

References

1. Brown, W. (1910). Some experimental results in the correlation of mental abilities. Brit. J. Psychol., 3(3), 296–322. Doi: http://dx.doi.org/10.1111/j.2044-8295.1910.tb00207.x
2. Callender, J., & Osburn, H. (1979). An Empirical Comparison of Coefficient Alpha, Guttman's Lambda - 2, and MSPLIT Maximized Split-Half Reliability Estimates. Journal of Educational Measurement, 16(2), 89–99. Doi: http://dx.doi.org/10.1111/j.1745-3984.1979.tb00090.x.
3. Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. Psychometrika, 16(3) Sept. 297–334. Doi: http://dx.doi.org/10.1007/BF02310555.
4. Ebel, R. L. (1967). The relation of Item Discrimination to test reliability. Journal of Educational Measurement, 4(3), 125–128. Doi: http://dx.doi.org/10.1111/j.1745-3984.1967.tb00579.x.
5. Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability. What They Are and How to Use Them. Educational and Psychological Measurement, 66(6), 930-944. Doi: http://dx.doi.org/10.1177/0013164406288165 Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?-doi=10.1.1.408.4091&rep=rep1&type=pdf (Accessed September 22, 2016).
6. Gulliksen, H. (1950). Theory of Mental Tests. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
7. Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10(4), 255-282. Doi: http://dx.doi.org/10.1007/BF02288892.
8. Guttman, L. (1950). The basis for scalogram analysis. In SA Stouffer, L Guttman, EA, Suchman, PF, Lazarsfield, SA Star, & JA Clausen (Eds.), Measurement and prediction. Princeton: Princeton University Press.
9. Henrysson, S. (1963). Correction of item-total correlations in item analysis. Psychometrika, 28(2), 211–218. Doi: http://dx.doi.org/10.1007/BF02289618.
10. Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. Educational and Psychological Measurement, 60(4), 523–531. Doi: http://dx.doi.org/10.1177/00131640021970691.
11. IBM. (2011). IBM SPSS Statistics 20 Algorithms. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf. (Accessed September 29, 2016).
12. Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. Psychometrika, 42(4), 567–578. Doi: http://dx.doi.org/10.1007/BF02295979.
13. Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30(1), 17–24. Doi: http://dx.doi.org/10.1037/h0057123. Retrieved from http://www.inf.ufsc.br/~cezar/tri_material/Artigo27%25.pdf. (Accessed April 30th, 2016)
14. Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. Psychometrika, 39(4), 491–499. Doi: http://dx.doi.org/10.1007/BF02291670.
15. Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160. Doi: http://dx.doi.org/10.1007/BF02288391.
16. Linacre, J. M. & Wright, B. D. (1996). Guttman-style item location maps. Rasch Measurement Transactions, 10(2), 492–493.
17. Lord, F. M. and Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, Mass: Addison –Wesley Publishing Company.
18. Metsämuuronen, J. (2017). Basics of Test Theory and Test Construction. In J. Metsämuuronen, Essentials of Contemporary Research Methods in Human Sciences. Vol 1: Elementary Basics, Section II. SAGE Publications. [Forthcoming]
19. Novick, M.R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32(1), 1–13. Doi: http://dx.doi.org/10.1007/BF02289400.
20. Raykov, T. (1997). Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence for Fixed Congeneric Components. Multivariate Behavioral Research, 32(4), 329-354. Doi: http://doi.org/10.1207/s15327906mbr3204_2.
21. Raykov, T., & Marcoulides, G. A. (2012). Evaluation of Validity and Reliability of Hierarchical Scales. Structural Equation Modeling, 19(3), 495-508. Doi: http://dx.doi.org/10.1080/10705511.2012.687675.
22. Raykov, T., & Marcoulides, G. A. (2013). Meta-Analysis of Reliability Coefficients Using Latent Variable Modeling. Structural Equation Modeling, 20(2), 338-353. Doi: http://dx.doi.org/10.1080/10705511.2013.769396.
23. Raykov, T., & Marcoulides, G. A. (2015). Scale Reliability Evaluation in Heterogeneous Populations. Educational and Psychological Measurement, 75(5), 875-892. Doi: http://dx.doi.org/10.1177/0013164414558587.
24. Raykov, T., & Marcoulides, G. A. (2016). Scale Reliability Evaluation Under Multiple Assumption Violations. Structural Equation Modeling, 23(2), 302-313. Doi: http://dx.doi.org/10.1080/10705511.2014.938597.
25. Raykov, T., & Traynor, A. (2016). Evaluation of Scale Reliability in Complex Sampling Designs. Structural Equation Modeling, 23(2), 270-277. http://dx.doi.org/10.1080/10705511.2014.938219.
26. Raykov, T., West, B. T., & Traynor, A. (2015). Evaluation of Coefficient Alpha for Multiple Component Measuring Instruments in Complex Sample Designs. Structural Equation Modeling, 22(3), 429-438. http://dx.doi.org/10.1080/10705511.2014.936081.
27. Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 9, 99–103.
28. Spearman, C. (1910). Correlation computed with faulty data. Brit. J. Psychol., 3(3), 271–295. Doi: http://dx.doi.org/10.1111/j.2044-8295.1910.tb00206.x.
29. Stanley, J. T. (1964). Measurement in today's schools. 4th ed. Englewood Cliffs, N.J.: Prentice-Hall.
30. Tarkkonen, L. (1987). On Reliability of Composite Scales. An Essay on the measurement and the properties of the coefficients of reliability - unified approach. Tilastotieteellisiä tutkimuksia 7. Finnish Statistical Society, Helsinki.
31. ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to reliability and minimum trace factor analysis. Psychometrika, 46(2), 201–213. Doi: http://dx.doi.org/10.1007/ BF02293900.
32. ten Berge, J. M. F, & Sočan, G (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika, 69(4), 613–625. Doi: http://dx.doi.org/10.1007/BF02289858.
33. ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. Psychometrika, 43(4), 575–579. Doi: http://dx.doi.org/ 10.1007/BF02293815.
34. Vehkalahti, K. (1995). Reliabiliteettimittojen tilastollisista ominaisuuksista. [Statistical properties of the estimators of reliability]. Tilastotieteen laitos, Helsinki: Helsingin yliopisto.
35. Vehkalahti, K. (2000). Reliability of Measurement Scales. Statistical Research Reports 17. Finnish Statistical Society. Retrieved from http://ethesis.helsinki.fi/julkaisut/val/tilas/vk/vehkalahti/ (Accessed 24.8.2016).
36. Yang, Y. & Green, S.B. (2011). Coefficient Alpha: A Reliability Coefficient for the 21st Century? Journal of Psychoeducational Assessment, 29(4), 377-392. Doi: http://dx.doi.org/10.1177/0734282911406668.