GENERALIZED RELIABILITY USED FOR THE COMPARISON OF RELIABILITY ESTIMATES IN TESTS WITH BINARY ITEMS

Patrícia Martinková and Marie Turčičová

Institute of Computer Science AS CR, Prague, Czech Republic

patricia.martinkova@gmail.com

1. Introduction

Reliability of measurement is a measure of its reproducibility under replicate conditions. The most widely used estimator of reliability is *Cronbach's alpha*. Nevertheless, there is an ongoing debate about the usefulness of alpha [1]. In tests composed of dichotomously scored items, the use of alpha is doubtful because it assumes that the item responses are continuous.

This paper presents incorporation of generalized linear models into a definition of reliability, which empowers the study of properties of various reliability estimates for tests with binary items.

2. Generalized definition of reliability

In the context of the classical test theory (CTT) the measurement Y is assumed to be composed out of two independent variables – the true value T and the error term ε

$$Y = T + \varepsilon, \quad T \sim (\mu, \sigma_T^2), \varepsilon \sim (0, \sigma^2). \tag{1}$$

The reliability of measurement Y can be defined as the ratio of variance of the true score and the variance of the observed score

$$reli(Y) = \frac{\mathrm{var}(T)}{\mathrm{var}(Y)} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}. \tag{2}$$

However, the model (1) is not appropriate when Y takes only values of 0 or 1 since in such a situation Y cannot be expressed as a sum of two independent variables. Instead, the model is usually defined through conditional mean values $E(Y|T)$.

One of the often used models is a generalized linear model of logistic regression, the so called **Rasch model**

$$\mathrm{E}(Y_{ij}|T_i) = \frac{\exp(T_i + \beta_j)}{1 + \exp(T_i + \beta_j)}, \quad i = 1, \ldots, n, j = 1, \ldots, m, \tag{3}$$

where n denotes the number of subjects who answered to m items with item difficulties $\beta_1, \ldots, \beta_m$.

Such models do not allow for rewriting the reliability as in (2). Hence, in [2] we proposed to use decomposition of $\mathrm{var}(Y)$ by means of conditional variance and conditional mean value as

$$\mathrm{var}(Y) = \mathrm{E}(\mathrm{var}(Y|T)) + \mathrm{var}(\mathrm{E}(Y|T)), \tag{4}$$

where the first term is the intraclass variance (that is the part of the variance, which is not due to the variability of T) and the second term is the interclass variance (the part of total variance which is due to the variability of T).

Using the *variance decomposition formula* (4) and following the CTT definition of reliability (2), we proposed a **generalized definition of reliability**

$$reli(Y) = \frac{\mathrm{var}[\mathrm{E}(Y|T)]}{\mathrm{var}(Y)} = \frac{\mathrm{var}[\mathrm{E}(Y|T)]}{\mathrm{E}[\mathrm{var}(Y|T)] + \mathrm{var}[\mathrm{E}(Y|T)]}. \tag{5}$$

Since for the CTT model (1) holds that

$$\mathrm{E}(Y|T) = \mathrm{E}(T + \varepsilon|T) = T,$$

in CTT, the definition (5) coincides with the classical definition (2).

In [3] we derived that for the composite measurement $Y_\bullet = \sum_{j=1}^m Y_j$ with essentially tau-equivalent items obeying the Rasch model, the reliability can be written as

$$reli(Y_\bullet) = \frac{\sum_{j=1}^m \sum_{t=1}^m (C_{jt} - D_j D_t)}{\sum_{j=1}^m \sum_{t=1}^m (C_{jt} - D_j D_t) + \sum_{j=1}^m B_j}, \tag{6}$$

where

$$B_j = E_T \frac{e^{T+\beta_j}}{\left(1 + e^{T+\beta_j}\right)^2}, \quad D_j = E_T \frac{e^{T+\beta_j}}{1 + e^{T+\beta_j}}$$

$$C_{jt} = E_T \frac{e^{T+\beta_j}}{1 + e^{T+\beta_j}} \frac{e^{T+\beta_t}}{1 + e^{T+\beta_t}}.$$

These integrals cannot be evaluated explicitly, nevertheless they can be evaluated numerically.

Hence, generalized definition (5), and formula (6) for the Rasch model, allow us to compute the true value of reliability for a given testing situation (given distribution of true values T, number of items m and item difficulties $\beta_1, \ldots, \beta_m$). Further, we are able to compare the properties of different reliability estimators in tests with binary items.

In the following sections we present examples of simulation studies empowered by the generalized definition (5).

3. Cronbach's alpha versus KR-21

The most widely used estimator of the reliability of composite measurement $Y_\bullet$ is **Cronbach's alpha**

$$\alpha = \frac{m}{m-1} \frac{\mathrm{var}(Y_\bullet) - \sum_{j=1}^m \mathrm{var}(Y_j)}{\mathrm{var}(Y_\bullet)}. \tag{7}$$

For binary data, the alpha coincides with **Kuder-Richardson formula 20**

$$KR\text{-}20 = \frac{m}{m-1} \left[1 - \frac{\sum_{j=1}^m \pi_j (1 - \pi_j)}{\mathrm{var}(Y_\bullet)} \right], \tag{8}$$

where π_j is the probability of a correct answer to item j.

For tests with equally difficult items, Kuder and Richardson developed **formula-21** [4], which is defined as

$$KR\text{-}21 = \frac{m}{m-1} \left[1 - \frac{\mathrm{E}(Y_\bullet)(m - \mathrm{E}(Y_\bullet))}{m \mathrm{var}(Y_\bullet)} \right]. \tag{9}$$

Assuming the Rasch model (3) and using a generalized definition of reliability (5), (6) we compared KR-20 and KR-21 in a simulation study. The simulations were conducted in $\mathbb{R}$, various input conditions were considered – various combinations of number of subjects n and number of essentially tau-equivalent items m. Moreover, we considered two different sets of item difficulties: $\beta_1, \ldots, \beta_m$ were chosen equidistant from interval $(-0.1, 0.1)$ or $(-3, 3)$.

In the case of small differences among the item difficulties ($\beta_j \in (-0.1, 0.1)$) and when the number of subjects was bigger than the number of items, Cronbach's alpha and KR-21 almost coincided, KR-21 gave slightly better results (lower bias) than alpha. This result is in accordance with [5].

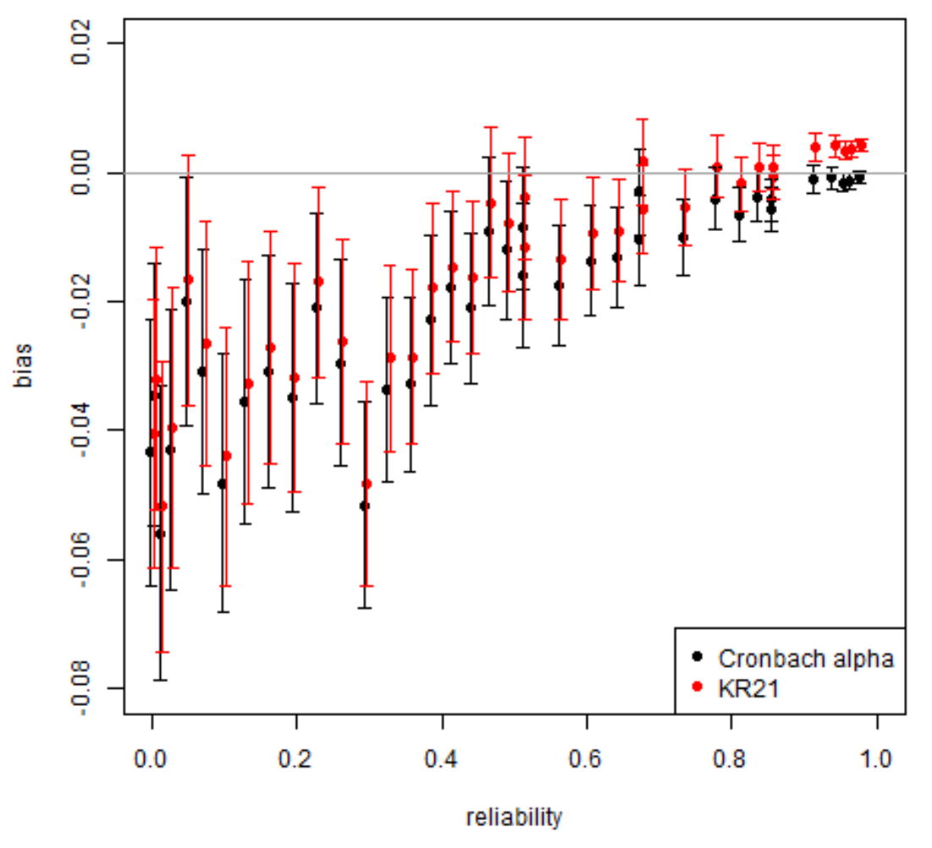

Fig. 1: Bias of two estimators of reliability, item difficulties from (-0.1,0.1). Number of students n=50, number of items m=5, number of simulations 500.

In the case of large differences among item difficulties ($\beta_j \in (-3, 3)$), the KR-21 provided more biased results than Cronbach's alpha. This is not surprising considering that KR-21 was proposed for equally difficult items.

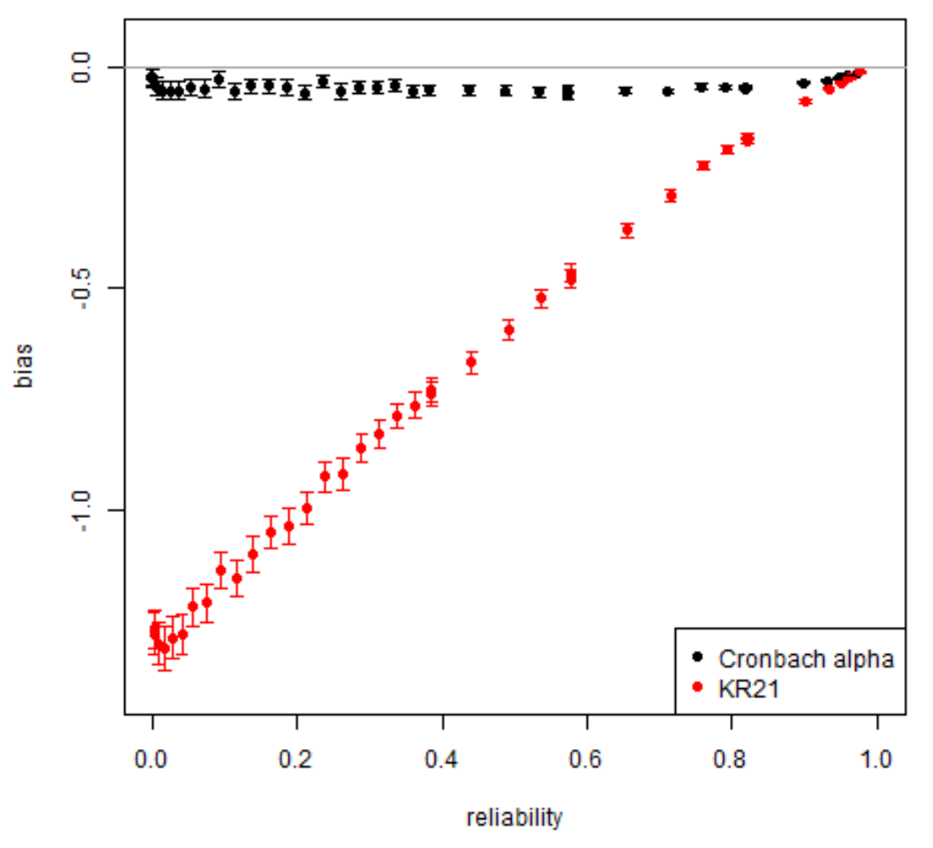

Fig. 2: Bias of two estimators of reliability, item difficulties from (-3,3). Number of students n=50, number of items m=5, number of simulations 500.

In accordance with theory, both estimators underestimated the true reliability.

4. Logistic alpha

The generalized definition of reliability (5) was originally proposed to study properties of the newly proposed estimate of reliability, the so called **logistic alpha** [3]. Motivation for it is as follows.

Considering 2-way ANOVA mixed effects model

$$Y_{ij} = T_i + \beta_j + \varepsilon_{ij}, \tag{10}$$

and adding the assumptions of normality ($T_i \sim N(\mu, \sigma_T^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$), the Cronbach's alpha (7) can be expressed as

$$\alpha = \frac{m\sigma_T^2}{m\sigma_T^2 + \sigma^2} = \frac{\mathrm{EMS}_A - \mathrm{EMS}_E}{\mathrm{EMS}_A}, \tag{11}$$

where

$$\mathrm{EMS}_A = m\sigma_T^2 + \sigma^2,$$
$$\mathrm{EMS}_E = \sigma^2$$

are the expectations of mean squares from ANOVA model. Hence, the Cronbach's alpha (11) can be estimated as

$$\hat{\alpha} = \frac{MS_A - MS_E}{MS_A} = 1 - \frac{MS_E}{MS_A} = 1 - \frac{1}{F_T}, \tag{12}$$

where F_T is the statistic used for testing the submodel with no subject effect i in the full model (10). The F_T statistic is best suited for normally distributed data. For dichotomous data we might think of replacing F_T by an analogous statistic from logistic regression. In the fixed effects model of logistic regression, the appropriate statistic is the difference of deviances in the submodel and in the model

$$X^2 = D(B) - D(A + B).$$

This statistic has under the null hypothesis asymptotically (for n fixed and m approaching infinity) the χ^2 distribution with $(n-1)$ degrees of freedom. Hence, the proposed estimate of reliability for composite dichotomous measurements, **logistic alpha** [3], is

$$\hat{\alpha}_{log} = 1 - \frac{n-1}{X^2}. \tag{13}$$

In the simulation study, logistic alpha had a lower bias than Cronbach's alpha for small differences between n and m and for true reliability not close to one. In all explored cases the curves of the bias of both estimates crossed in the right side of the graph.

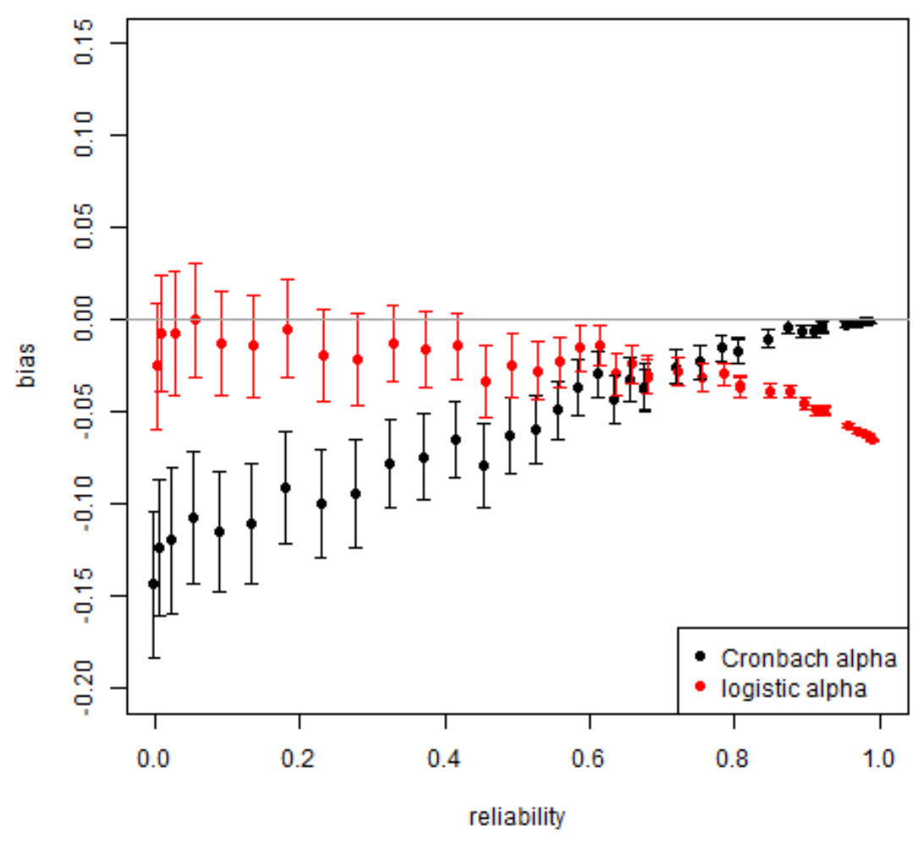

Fig. 3: Bias for two estimates of reliability, item difficulties from (-0.1,0.1). Number of students n=20, number of items m=10, number of simulations 500.

Inferior results of the logistic alpha were obtained for true reliability close to 1 and when the number of students was high in proportion to the number of items.

5. Conclusion and discussion

We proposed the generalized definition of reliability which coincides with classical definition in the CTT and moreover is appropriate for tests composed of binary items.

We demonstrated usage of generalized reliability for comparison of different reliability estimators: Cronbach's alpha, KR-21 and the newly proposed estimate *logistic alpha*. We demonstrated that KR-21 is not appropriate for the items with unequal difficulties. The new estimate logistic alpha provided interesting results and it would be worthwhile to explore the possibilities of its correction with the aim to lower the bias also for reliability close to 1.

The concept can be used to study properties of various reliability estimates in tests composed of binary items. In this paper, items were assumed to be essentially tau-equivalent, further research might concentrate on more complicated testing schemes.

References

[1] Sijtsma K. On the use, the misuse and very limited usefulness of Cronbachs alpha. *Psychometrika*, 74(1), pp. 107-120, 2009.

[2] Martinková P, Zvára K. Reliability of Composite Dichotomous Measurements. *European Journal for Biomedical Informatics*, 6(2), pp. 14-23, 2010.

[3] Martinková P, Zvára K. Reliability in the Rasch model. *Kybernetika*, 43(3), pp. 315-326, 2007.

[4] Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika*, 2(3), pp. 151-160, 1937.

[5] Brennan RL (Ed.) *Educational measurement.* Praeger Publishers, 4th edition, p. 100, 2006.

IMPS 2013, July 22.-26., 2013, Arnhem, the Netherlands