Reliability of Composite Dichotomous Measurements

Patrícia Martinková[1], **Karel Zvára**[2]

[1] Centre of Biomedical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic,

[2] Department of Medical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

Summary

Reliability of measurement is a measure of its reproducibility under replicate conditions. The classical concept of reliability assumes that measurement *Y* is composed out of true value *T* and error term ε, two independent random variables, *Y* = *T* + ε . Reliability of measurement is defined as the ratio of the variance of the true scores to the variance of the observed scores. However, this concept is not applicable in models for dichotomous measurements which do not consider error terms and are instead defined via conditional probabilities. In this paper we examine a more general definition of reliability proposed in [1], which is based on decomposition of variance in mixed effects model. Proposed definition covers the classical definition of reliability and it is, moreover, appropriate for dichotomous measurements, too. Newly, for the proposed definition assumptions are derived, under which the reliability of composite measurement can be predicted by reliability of single measurement (Spearman-Brown formula) and approximate validity of Spearman-Brown formula is shown for the Rasch model. Finally, as a modification of the classical estimate of reliability based on Cronbach's alpha, we examine its counterpart *logistic alpha* introduced in [2], which appears to be more appropriate for composite dichotomous measurements in some cases. Simulations show that the new estimate does not tend to underestimate reliability as often as the Cronbach's alpha does. The new estimate is used in binary data of computerized process of myocardial perfusion diagnosis from cardiac single proton emission computed tomography (SPECT).

Keywords: reliability, binary data, logistic regression, Cronbach alpha, Rasch model, myocardial perfusion diagnosis

1. Introduction and statistical background

Reliability of measurement is a measure of its reproducibility under replicate conditions. In medical practice, the reliability of measurement remains an important topic engendering much discussion. For continuous measurements, reliability analysis and equivalence test for agreement were lately studied by Yi, Wang and He [3]. A nonparametric, probabilistic estimate of reliability used on cognitive tests in Alzheimer's disease was examined by van Belle and Arnold [4]. IRT model-based reliability estimates, which are appropriate for dichotomous or ordinal outcomes, were used in Teresi et. al [5].

Reproducibility studies for binary outcomes are typically analysed using kappa statistics, which was motivated by its relation to the intraclass correlation coefficient [6], [7]. In this paper we take quite a different approach – we discuss the decomposition of variance in mixed effects model settings, as appeared in [8] and propose a new definition of reliability, which covers the classical testing situation and is moreover suitable also for binary data. We also discuss a new estimate of reliability. For better understanding of parallel with classical test theory, a summary of the basic principles of the classical test theory (CTT) is given in this section.

1.1 Reliability of measurement within CTT

In the classical test theory [9], it is assumed that the measurement *Y* is composed out of the true value *T* and the error term ε, independent continuous random variables

$$Y = T + \epsilon,$$

$$T \sim (\mu, \sigma_T^2),\ \sigma_T^2 > 0,$$

$$\epsilon \sim (0, \sigma^2),\ \sigma^2 > 0. \qquad (1)$$

The reliability of measurement is defined as a ratio of variance of the true score and variance of the observed score

$$reli(Y) = \frac{\mathrm{var}\,(T)}{\mathrm{var}\,(Y)} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2} = \rho_1. \qquad (2)$$

Alternatively, reliability can be defined as a squared correlation between the measured value and the measurement

$$\mathrm{corr}^2(Y,T) = \mathrm{corr}^2(T+\epsilon, T) = \frac{(\sigma_T^2)^2}{(\sigma_T^2+\sigma^2)\sigma_T^2} = \mathrm{reli}(Y). \qquad (3)$$

Also, reliability can be expressed as a correlation between repeated measurements Y_1, Y_2, that is between two independent and equally accurate measurements of the same true value *T*

$$Y_j = T + \epsilon_j,\ j = 1, 2,$$

$$\epsilon_1, \epsilon_2 \sim (0, \sigma^2) \text{ independent}$$

$$\mathrm{corr}\,(Y_1, Y_2) = \mathrm{corr}\,(T + \epsilon_1, T + \epsilon_2) =$$

$$= \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2} = \mathrm{reli}(Y) \qquad (4)$$

Correlation between two independent (not necessary equally accurate) measurements of the same true value can be expressed as

$$Y_j = T + \epsilon_j,\ j = 1, 2,$$

$$\epsilon_1 \sim (0, \sigma_1^2),\ \epsilon_2 \sim (0, \sigma_2^2) \text{ independent}$$

$$\mathrm{corr}\,(Y_1, Y_2) = \mathrm{corr}\,(T + \epsilon_1, T + \epsilon_2) =$$

$$= \frac{\sigma_T^2}{\sqrt{\sigma_T^2 + \sigma_1^2}\sqrt{\sigma_T^2 + \sigma_2^2}} = \sqrt{\mathrm{reli}(Y_1)}\sqrt{\mathrm{reli}(Y_2)} \qquad (5)$$

1.2 Reliability of sum of repeated measurements

Having *J* repeated measurements of *T*

$$Y_j = T + \epsilon_j, \quad \epsilon_j \sim (0, \sigma^2)\, iid, \quad j = 1, \ldots, J, \qquad (6)$$

variability of their sum $Y_\bullet = \sum_j Y_j$ is

$$\mathrm{var}\,(Y_\bullet) = \mathrm{var}\left(\sum_j (T + \epsilon_j)\right) = J^2\sigma_T^2 + J\sigma^2, \tag{7}$$

hence reliability p_J of sum $Y_\bullet$ or of average

$$\bar{Y}_\bullet = \frac{1}{J} Y_\bullet$$

can be expressed by means of reliability of single measurement p_1 (Spearman-Brown formula, see [10], [11]).

$$\mathrm{reli}(Y_\bullet) = \frac{\mathrm{var}\,(JT)}{\mathrm{var}\,(Y_\bullet)} = \frac{J^2\sigma_T^2}{J^2\sigma_T^2 + J\sigma^2} =$$

$$= \frac{J\rho_1}{(J-1)\rho_1 + 1} = \rho_J. \tag{8}$$

Reliability of sum of J repeated measurements can also be expressed in a more practical way

$$\mathrm{reli}(Y_\bullet) \quad = \quad \frac{\mathrm{var}\,(JT)}{\mathrm{var}\,(Y_\bullet)} = \frac{J}{J-1}\,\frac{J-1}{J}\,\frac{J^2\sigma_T^2}{\mathrm{var}\,(Y_\bullet)} =$$

$$= \frac{J}{J-1}\,\frac{J^2\sigma_T^2 + J\sigma^2 - J\sigma^2 - J\sigma_T^2}{\mathrm{var}\,(Y_\bullet)}$$

$$= \frac{J}{J-1}\,\frac{\mathrm{var}\left(\sum_j Y_j\right) - \sum_j \mathrm{var}\,(Y_j)}{\mathrm{var}\left(\sum_j Y_j\right)} =$$

$$= \frac{J}{J-1}\,\frac{\sum\sum_{j\neq k}\sigma_{jk}}{\sum\sum_{j,k}\sigma_{jk}} = \alpha \tag{9}$$

where $\sigma_{jk} = \mathrm{cov}(Y_j, Y_k)$. In equation (9), we got so called **Cronbach's alpha** (see [12]) which may be estimated by using sample covariances instead of their population counterparts

$$\hat{\alpha} = \frac{J}{J-1}\,\frac{\sum\sum_{j\neq k} s_{jk}}{\sum\sum_{j,k} s_{jk}},$$

$$\text{kde } s_{jk} = \frac{1}{I-1}\sum_{t=1}^{I}(Y_{tj} - \bar{Y}_{\bullet j})(Y_{tk} - \bar{Y}_{\bullet k}). \tag{10}$$

For dichotomous data, estimate (10) coincides with Kuder-Richardson formula 20 (see [13]).

As shown above in (9), Cronbach's alpha is equivalent to reliability of sum of repeated measurements of the *same* true value T. Nevertheless, it is also widely used as an estimator of reliability of composite measurements.

1.3 Reliability of composite measurements

Often, the measurement cannot be repeated independently to produce exactly the same true value T. In psychometrics, the tests are composed of J items where each of concentrates on a slightly different aspect of measured quantity T_j; each subject is described by the sum of J item scores $Y_\bullet = \sum Y_j$.

In medical practice, the health professionals are often faced with the same quantitative measurements reported by different raters, or from the same rater measured using different tools, and the measured property is often described by the average of the measurements

$$\bar{Y}_\bullet = \sum Y_j / J$$

For j-th measurement, we suppose that

$$Y_j - T_j + \epsilon_j, \qquad \epsilon_j \sim (0, \sigma_j),$$

$$j - 1, \ldots, J \tag{11}$$

where $(\varepsilon_1,\ldots,\varepsilon_J)$ are mutually independent and also independent of $(T_1,\ldots,T_J)$ Reliability of composite measurement

$$Y_\bullet = \sum_j Y_j$$

is

$$\mathrm{reli}(Y_\bullet) = \frac{\mathrm{var}\,(T_\bullet)}{\mathrm{var}\,(T_\bullet) + \mathrm{var}\,(\epsilon_\bullet)}. \tag{12}$$

As demonstrated by Novick and Lewis in [14], Cronbach's alpha is generally a lower bound of reliability of composite measurement

$$\alpha \leq \mathrm{reli}(Y_\bullet) = \rho_J. \tag{13}$$

Novick and Lewis showed that equality holds only for essentially τ-equivalent items, that is in case, where for a random variable T and real numbers β_j such that

$$\sum \beta_j = 0,$$

with probability = 1 holds

$$T_j = T + \beta_j, \quad \forall j = 1,\ldots,J. \tag{14}$$

This is equivalent to simultaneously holding

$$\mathrm{var}\,(T_j) = \sigma_T^2, \quad \forall j \tag{15}$$

$$\mathrm{corr}\,(T_j, T_k) = 1, \quad \forall j, k. \tag{16}$$

Condition (1.3) is needed to have equal item reliabilities in model (11). For Spearman-Brown formula to hold, also condition (16) is needed.

When items of composite measurement are not essentially τ-equivalent, besides the fact, that Spearman-Brown formula does not hold, from (13) we might also expect that estimate of reliability based on Cronbach's alpha will underestimate the true reliability. Some estimations of this discrepancy between reliability of composite measurement and Cronbach's alpha on population level can be found in [15].

1.4 Cronbach's alpha as estimator of reliability of composite measurements

Let us now suppose measurements Y_{ij}, $j = 1,\ldots,J$ on subjects $i = 1,\ldots,I$. Assumptions of essential τ-equivalence lead to 2-way mixed effects ANOVA model.

$$Y_{ij} = T_i + \beta_j + \epsilon_{ij}, \qquad \sum_j \beta_j = 0,$$

$$i = 1,\ldots,I, \quad j = 1,\ldots,J \tag{17}$$

where β_j is the item parameter (item difficulty, expert's level, etc.). When we add assumptions of normality

$$T_i \sim \mathrm{N}(\mu, \sigma_T^2), \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2),$$

and consider sum of squares decomposition

$$SS_T = \sum (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = SS_A + SS_B + SS_E,$$

the mean squares MS_A and MS_E have the following expectations

$$\mathrm{E}\, MS_A = \mathrm{E} \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{Y}_{i\bullet} - Y_{\bullet\bullet})^2/(I-1) = J\sigma_T^2 + \sigma^2, \tag{18}$$

$$\mathrm{E}\, MS_E = \mathrm{E} \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} \quad \bar{Y}_{i\bullet} \quad \bar{Y}_{\bullet j} \mid \bar{Y}_{\bullet\bullet})^2/$$

$$/((I-1)(J-1)) = \sigma^2 \tag{19}$$

Hence, Cronbach's alpha (9) can be expressed as

$$\alpha = \frac{\mathrm{E}\, MS_A - \mathrm{E}\, MS_E}{\mathrm{E}\, MS_A} \tag{20}$$

and the estimate (10) can be rewritten as

$$\hat{\alpha} = \frac{MS_A - MS_E}{MS_A} = 1 - \frac{MS_E}{MS_A} = 1 - \frac{1}{F_T}, \tag{21}$$

where F_T is the statistic used for submodel testing with no subject effect in the full model (17). The same statistic is used also for testing the submodel in 2-way ANOVA fixed effects model [16].

From (21) we can conclude that high values of $\hat{\alpha}$ indicate that the composite measurement can distinguish between the subjects well. Hence, we may expect higher values of $\hat{\alpha}$ when the subjects' variability is high.

2. New developments for dichotomous measurements

When measurements *Y* takes only the values 0 or 1, the classical model described in part 1.1 is not appropriate anymore, since *Y* cannot be expressed as a sum of two independent random variables. Instead, the model should be defined through conditional mean values *E*(*Y*|*T*) Let us suppose a more general model

$$Y_{ij} \sim f(\cdot, T_i), \qquad T_i \sim iid,$$

$$(Y_{ij}|T_i),\ (Y_{ij'}|T_i) \text{ independent for } j \neq j \tag{22}$$

One of such models is the Rasch model (see [17]):

$$\mathrm{E}\,(Y_{ij}|T_i) = \pi(T_i, \beta_j) = \frac{\exp(T_i + \beta_j)}{1 + \exp(T_i + \beta_j)},$$

$$T_i \sim \mathrm{N}(\mu, \sigma_T^2). \tag{23}$$

In the framework of Item Response Theory (IRT) the Rasch model (23) and its generalizations are widely studied [18], especially in connection with parameter estimation. The concept of reliability is extended from a single index to a function of the true value *T* called the *test information function* [19]. Besides, it is also possible to obtain an index for a test as a whole which is directly analogous to Cronbach's alpha: As an analogy to decomposing an observed score into true score and an error in classical test theory, we consider a decomposition of an IRT person estimate into a true location and error

$$\hat{T} = T + \epsilon. \tag{24}$$

The reliability in IRT is defined as

$$R_{IRT} = \frac{\mathrm{var}\,(T)}{\mathrm{var}\,(\hat{T})} = \frac{\mathrm{var}\,(\hat{T}) - \mathrm{var}\,(\epsilon)}{\mathrm{var}\,(\hat{T})}. \tag{25}$$

For its estimation, first, the estimates of subjects' true locations and their standard errors (SE) are computed by standard estimation procedures. Then, the sample variance of these estimates is computed to estimate var($\hat{T}$) The mean squared subject standard error estimate provides an estimate of the variance of the error *var*(ε). The resulting estimate is typically very close to Cronbach's alpha [20].

In this paper, we consider a different definition of reliability for model (22), one that is a more straightforward generalization of classical reliability. We give formula for reliability in the Rasch model and assumptions for Spearman-Brown formula to hold in model (22). We also introduce a new estimate of reliability appropriate for composite dichotomous measurements and we compare it to the Cronbach's alpha in simulations and in practical example.

2.1 Proposed definition of reliability

The total observed variance var(Y_{ij}) can be decomposed by the means of conditional variance and conditional mean value as

$$\mathrm{var}\,(Y_{ij}) = \mathrm{E}\,(\mathrm{var}\,(Y_{ij}|T_i)) + \mathrm{var}\,(\mathrm{E}\,(Y_{ij}|T_i)). \tag{26}$$

where the first term is the intraclass variance, that is the part of the variance, which is not due to the variability of T_i and the second term is the interclass variance, the part of total variance which is due to the variability of T_i [8].

To follow the definition of reliability from the classical test theory, we might define it as a ratio of the variance due to variability of the measured property *T* to the total observed variability, that is

$$\mathrm{reli}(Y) = \frac{\mathrm{var}\,(\mathrm{E}\,(Y|T))}{\mathrm{var}\,(Y)} = \frac{\mathrm{var}\,(\mathrm{E}\,(Y|T))}{\mathrm{var}\,(\mathrm{E}\,(Y|T)) + \mathrm{E}\,(\mathrm{var}\,(Y|T))}. \tag{27}$$

Since for the classical model holds

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

definition (27) coincides with the classical definition (2).

For composite measurements, the reliability of *j*-th item can be defined as

$$\mathrm{reli}(Y_{ij}) = \frac{\mathrm{var}\,(\mathrm{E}\,(Y_{ij}|T_i))}{\mathrm{var}\,(Y_{ij})} = \tau_{ij} \tag{28}$$

and the reliability of composite measurement $Y_{i\bullet}$ can be defined as

$$\mathrm{reli}(Y_{i\bullet}) = \frac{\mathrm{var}\,(\mathrm{E}\,(Y_{i\bullet}|T_i))}{\mathrm{var}\,(Y_{i\bullet})}. \tag{29}$$

For the Rasch model (23), we derived in [1] that reliability of composite measurement is

$$\mathrm{reli}(Y_\bullet) = \frac{\mathrm{var}\,(\mathrm{E}\,(Y_\bullet|T))}{\mathrm{var}\,(Y_\bullet)} = \frac{\sum_{j=1}^{J}\sum_{t=1}^{J}(C_{jt} - D_jD_t)}{\sum_{j=1}^{J}\sum_{t=1}^{J}(C_{jt} - D_jD_t) + \sum_{j=1}^{J} B_j}, \tag{30}$$

where

$$B_j = \mathrm{E}_T \frac{e^{T+\beta_j}}{(1+e^{T+\beta_j})^2},$$

$$D_j = \mathrm{E}_T \frac{e^{T+\beta_j}}{1+e^{T+\beta_j}},$$

$$C_{jl} = \mathrm{E}_T \frac{e^{T+\beta_j}}{1+e^{T+\beta_j}} \frac{e^{T+\beta_l}}{1+e^{T+\beta_l}}$$

These integrals cannot be evaluated explicitly, nevertheless they can be evaluated numerically. Hence, for a given testing situation (that is for distribution of subjects' true values *T* number of items *J* and their levels β*ⱼ*, *j* = 1,...,*J*,) the true value of reliability can be computed (see Table 1).

2.2 Spearman-Brown formula

Preliminary let's find assumptions for having equal item reliabilities τ_{ij} for all *j*. As we have already mentioned, reliability can also be expressed as a correlation of two independent measurements of the same property *T* (see formula (4)), that is by the (i-th subject) intraclass correlation

. $$\rho_{ijj'} = \mathrm{corr}\,(Y_{ij} Y_{ij'}).$$

Within model (22), the relationship between $P_{ijj'}$ and τ_{ij}, $\tau_{ij'}$ for j ≠ j´ is following (for proof, see [21]):

$$\rho_{ijj'} = \sqrt{\tau_{ij}}\sqrt{\tau_{ij'}}\,\mathrm{corr}\,[\mathrm{E}\,(Y_{ij}|T_i),$$

$$\mathrm{E}\,(Y_{ij'}|T_i)] \le \sqrt{\tau_{ij}}\sqrt{\tau_{ij'}} \qquad (31)$$

The equality in (31) holds for all j≠j´ if for all j≠j´

$$\mathrm{corr}\,[\mathrm{E}\,(Y_{ij}|T_i), \mathrm{E}\,(Y_{ij'}|T_i)] = 1, \qquad (32)$$

that is, in the case when for all *i*, *j* with probability equal to one, for some constants $k_{ij} > 0$ and η_{ij} and some functions $\lambda_i(T_i)$ the conditional means can be expressed as

$$\mathrm{E}\,(Y_{ij}|T_i) = k_{ij}\,[\lambda_i(T_i) + \eta_{ij}]\,. \qquad (33)$$

Moreover, assumption (33) can be required with additional constraint

$$\textstyle\sum_{k=1}^{J} k_{ij}^2 = J,$$

since λ_i and η_{ij} can be multiplied by appropriate constants.

Formula (31) and assumption (33) may be extended to the following theorem (for proof, see [21]), which revises the theorem of Commenges and Jacqmin (see [8]):

Theorem Suppose that Y_{ij} for *i* = 1,...,*I*, *j* = 1,...,*J*, *J* ≥ 3 obey the model (22). Moreover, with probability equal to one let for all *i*, *j* hold the assumption (33). Then the following propositions are equivalent:

P1 $p_{ijj'} = p_i$ does not depend on *j*, *j´* for any *j≠j´*.

P2 $\tau_{ij} = \tau_i$ does not depend on *j* for any, *j*.

P3 The model belongs to a class specified (with probability equal to one) by:

$$\mathrm{var}\,(Y_{ij}|T_i) = k_{ij}^2\,\left[\sigma_i^2(T_i) + \psi_{ij}(T_i)\right], \qquad (34)$$

where $\mathrm{E}\,[\psi_{ij}(T_i)] = 0$, and $\psi_{ij}(T_i) > -\sigma_i^2(T_i)$.

P4 $\rho_i = \tau_i$.

Hence, when assumptions (33) and (34) hold, the item reliabilities equal τ_i for all *j*. Let us now look at reliability of composite measurements under these two assumptions and under mentioned constraint

$$\textstyle\sum_{j=1}^{J} k_{ij}^2 = J.$$

The reliability of (every) single item may be written as

$$R_1 = \tau_{ij} = \frac{\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|T_i)]}{\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|T_i)] + \mathrm{E}\,[\mathrm{var}\,(Y_{ij}|T_i)]} =$$

$$= \frac{\mathrm{var}\,[\lambda_i(T_i)]}{\mathrm{var}\,[\lambda_i(T_i)] + \mathrm{E}\,[\sigma_i^2(T_i)]} = \tau_i = \rho$$

and by Theorem it coincides with the correlation of two independent measurements of the same property Y_{ij}, $Y_{ij'}$, j ≠ j´. The reliability of the composite measurement is

$$R_J = \frac{\mathrm{var}\,[\mathrm{E}\,(Y_i|T_i)]}{\mathrm{var}\,[\mathrm{E}\,(Y_i|T_i)] + \mathrm{E}\,[\mathrm{var}\,(Y_i|T_i)]} =$$

$$= \frac{\mathrm{var}\left[\sum_{j=1}^{J} \mathrm{E}\,(Y_{ij}|T_i)\right]}{\mathrm{var}\left[\sum_{j=1}^{J} \mathrm{E}\,(Y_{ij}|T_i)\right] + \sum_{j=1}^{J} \mathrm{E}\,[\mathrm{var}\,(Y_{ij}|T_i)]}$$

$$= \frac{\mathrm{var}\left\{\sum_{j=1}^{J} k_{ij}[\lambda_i(T_i) + \eta_{ij}]\right\}}{\mathrm{var}\left\{\sum_{j=1}^{J} k_{ij}[\lambda_i(T_i) + \eta_{ij}]\right\} + \sum_{j=1}^{J} k_{ij}^2 \mathrm{E}\,[\sigma_i^2(T_i) + \psi_{ij}(T_i)]}$$

$$= \frac{\left(\sum_{j=1}^{J} k_{ij}\right)^2 \mathrm{var}\,[\lambda_i(T_i)]}{\left(\sum_{j=1}^{J} k_{ij}\right)^2 \mathrm{var}\,[\lambda_i(T_i)] + m\mathrm{E}\,[\sigma_i^2(T_i)]} =$$

$$= \frac{\frac{\left(\sum_{j=1}^{J} k_{ij}\right)^2}{J} R_1}{1 + \left(\frac{\left(\sum_{j=1}^{J} k_{ij}\right)^2}{J} - 1\right) R_1}. \qquad (35)$$

The expression (35) coincides with Spearman-Brown formula if $k_{ij} = 1$ for all *j*. We may conclude that the assumptions (15)(16) of essential τ-equivalence for classical model correspond in model (22) with assumption that with probability equal to one, the conditional mean and variance of Y_{ij} may be written as

$$\mathrm{E}\,(Y_{ij}|T_i) = \lambda_i(T_i) + \eta_{ij}, \qquad (36)$$

$$\mathrm{var}\,(Y_{ij}|T_i) = \sigma_i^2(T_i) + \psi_{ij}(T_i), \qquad (37)$$

where η_{ij} are given constants,

$$\mathrm{E}\,[\psi_{ij}(T_i)] = 0,$$

and $\psi_{ij}(T_i) > -\sigma_i^2(T_i)$.

As showed in [21], the Rasch model (23) does not follow assumption (36) nor (37). Nevertheless, Table 1 and Table 2 give us an impression that the Spearman-Brown formula (8) does hold at least approximately.

In Table 1, the values of reliability were calculated from formula (30) for different testing situations (number of items *J* item levels equidistant on <-0.1, 0.1> subject levels $T_i \sim \mathrm{N}(0, \sigma_T^2)$).
To evaluate the integrals, function integrate in software R (see [22]) was used.

The maximum absolute error reached in integrations was less than 0.000025. In Table 2, we set $J = 11$ and used the second line of Table 1 together with the Spearman-Brown formula (8) to get approximate values of reliabilities for $J = 3, 20, 50$ and 100.

As we may see, the numerical values in Tables 1 and 2 are very similar. As an explanation we give an approximation of Spearman-Brown formula for the Rasch model. Let us assume b_j small,

$$\textstyle\sum_j b_j = 0,$$

and apply the first-order Taylor series approximation (to function of one or two variables)

$$B_j \;=\; \mathrm{E}\frac{e^{T+b_j}}{(1+e^{T+b_j})^2} \approx \mathrm{E}\frac{e^T}{(1+e^T)^2} +$$

$$+\, b_j \mathrm{E}\frac{e^T(1-e^T)}{(1+e^T)^3} = B + b_j \mathrm{E}\frac{e^T(1-e^T)}{(1+e^T)^3},$$

$$D_j \;=\; \mathrm{E}\frac{e^{T+b_j}}{1+e^{T+b_j}} \approx \mathrm{E}\frac{e^T}{1+e^T} +$$

$$+\, b_j \mathrm{E}\frac{e^T}{(1+e^T)^2} = D + b_j \mathrm{E}\frac{e^T}{(1+e^T)^2},$$

$$C_{jt} \;=\; \mathrm{E}\frac{e^{T+b_j}}{1+e^{T+b_j}}\frac{e^{T+b_t}}{1+e^{T+b_t}} \approx$$

$$\approx \mathrm{E}\frac{e^{2T}}{(1+e^T)^2} + (b_j+b_t)\mathrm{E}\frac{e^{2T}}{(1+e^T)^3} =$$

$$= C + (b_j+b_t)\mathrm{E}\frac{e^{2T}}{(1+e^T)^3}.$$

Then, the reliability of the composite measurement in the Rasch model is approximately

$$R_J \approx \frac{J^2(C-D^2)}{J^2(C-D^2)+JB}.$$

Tab. 1. Reliability in the Rasch model for different number of items.

Number	Variability of subjects σ_T^2						
of items	0.01	0.1	0.2	0.5	0.9	2.5	10
J=3	0.00008	0.00741	0.02881	0.15047	0.34335	0.73121	0.94152
J=11	**0.00028**	**0.02667**	**0.09814**	**0.39386**	**0.65731**	**0.90890**	**0.98335**
J=20	0.00050	0.04747	0.16519	0.54160	0.77717	0.94775	0.99077
J=50	0.00125	0.11078	0.33098	0.74709	0.89711	0.97843	0.99629
J=100	0.00249	0.19947	0.49735	0.85524	0.94577	0.98910	0.99814

Tab. 2. Spearman-Brown formula used for J=11.

Number	Variability of subjects σ_T^2						
of items	0.01	0.1	0.2	0.5	0.9	2.5	10
SB R_3	0.00008	0.00742	0.02882	0.15054	0.34345	0.73125	0.94153
SB R_{20}	0.00050	0.04746	0.16518	0.54159	0.77716	0.94775	0.99077
SB R_{50}	0.00125	0.11077	0.33095	0.74707	0.89710	0.97843	0.99629
SB R_{100}	0.00249	0.19944	0.49731	0.85522	0.94576	0.98910	0.99814

While the reliability of single measurement is

$$R_1 \approx \frac{C-D^2}{C-D^2+B},$$

which together gives an approximate validity of the Spearman-Brown formula in the Rasch model.

2.3 Estimate of reliability logistic alpha

F_T statistic in estimate based on Cronbach's alpha (21) is best suited for normally distributed data. For dichotomous data we might think of replacing F_T by 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 I fixed and approaching infinity) the χ^2 distribution with χ^2 degrees of freedom. Hence, the proposed estimate of reliability for composite dichotomous measurements, *logistic alpha* [2], [1] is:

$$\hat{\alpha}_{log} = 1 - \frac{I-1}{X^2}. \qquad (38)$$

In the following section we compare the new estimate *logistic alpha* to the classical estimate based on Cronbach's alpha.

3. Simulation example

This simulation is dedicated to the following example: $I = 20$ patients answered to $J = 20$ yes/no items of quality of life survey. The item levels β_j were supposed to be equidistant on <-0.1, 0.1>, patients' true values of QOL were assumed to be normally distributed with mean $\mu = 0$ and variance σ_T^2 (55 values of σ_T^2 were chosen from interval <0.01, 10> to get 55 values of reliability approximately uniformly distributed on interval).

For each combination of I, J and the true reliability was enumerated by formula (30) and 500 data sets generated: Set of I patients' life quality levels T_i was generated from $N(0, \sigma_T^2)$. QOL survey answers Y_{ij} were generated from the Rasch model (23) and estimates $\hat{\alpha}_{CR}$ and α_{log} were computed from the data.

From obtained 500 estimates $\hat{\alpha}_{CR}$ and 500 estimates $\hat{\alpha}_{log}$ the bias and mean squared error (MSE) were computed and plotted out in Figure 1 and 2.

Other testing situations (number of items J = 11 and I = 30 and number of patients I = 30 and 50) were studied, too. We observed smaller bias and MSE in $\hat{\alpha}_{log}$ particularly for true reliability ≤ 0.75. Inferior results of the *logistic alpha* were obtained for reliability close to and for high number of patients in proportion to the number of items. The latter might be a consequence of the fact, that while statistic X^2 used in (38) is appropriate for fixed effects model of logistic regression, in (23) we expect a mixed effects model. A chance of improvement lies in replacing statistic X^2 by even more appropriate one.

4. Analysis of cardiac data

The dataset **SPECT heart data** [23] describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 22 binary feature patterns (22 partial diagnoses based on SPECT) were created for each patient [24]. We were interested in internal consistency of the 22 partial diagnoses based on SPECT. Reached estimates of reliability were: Cronbach alpha $\hat{\alpha}_{CR}$ = 0.839 and logistic alpha $\hat{\alpha}_{log}$ = 0.827. In this case, both estimates are quite similar and show on high internal consistency of partial diagnoses.

5. Conclusion and discussion

In this paper, the basic principles from classical test theory (CTT) were summarized and used for a new definition of reliability and new estimate of reliability appropriate for composite dichotomous measurements. For classical testing situation the proposed definition of relibility was shown to coincide with definition of reliability in CTT. Assumptions for Spearman-Brown formula were given for model (22) which is more general than classical model (1). The proposed definition and estimate of reliability were applied in the Rasch model, for which the Spearman-Brown formula was shown to hold only approximately.

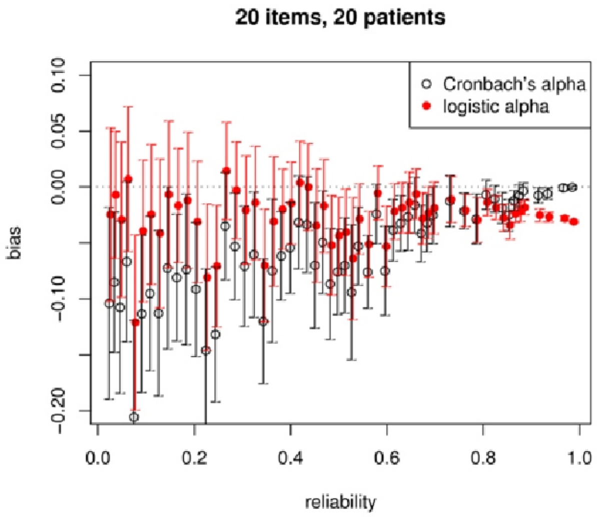

Fig. 1: Estimated Bias and its confidence interval for classical and logistic estimator of reliability

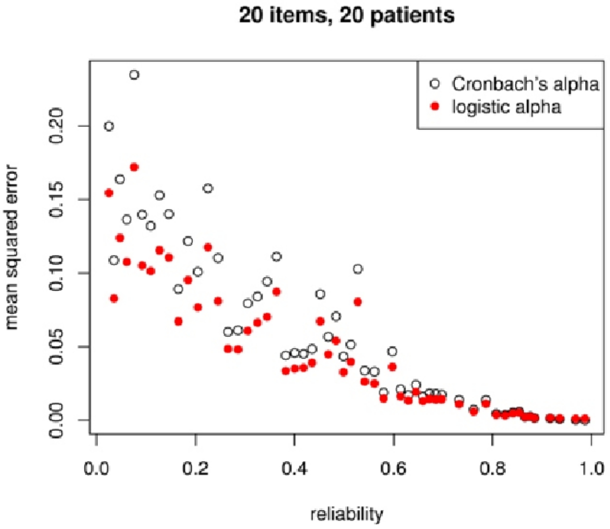

Fig. 2: Estimated MSE for classical and logistic estimator of reliability

The proposed estimate *logistic alpha* was shown to possess better properties (smaller bias and MSE), in particular for true reliability ≤ 0.75 and the number of items exceeding the number of patients. The chances of improvement of the new estimate for true reliability close to and for higher number of patients were discussed. Estimation of reliability on binary data was demonstrated on cardiac data. Work presented in this paper could lead to more precise estimation of reliability for binary data, which could contribute to many fields of biomedical research.

References

[1] Martinková P, Zvára K. Reliability in the Rasch model. Kybernetika 2007; 43(3):315326.

[2] Zvára K. Measuring of reliability: Beware of Cronbach. [Měření reliability aneb bacha na Cronbacha, in Czech]. Information Bulletin of the Czech Statistical Society 2002; 12:1320.

[3] Yi Q, Wang PP, He Y. Reliability analysis for continuous measurements: Equivalence test for agreement. Statistics in Medicine 2008; 27:28162825.

[4] vanBelle G, Arnold A. Reliability of cognitive tests used in Alzheimer's disease. Statistics in Medicine 2000; 19:14111420.

[5] Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Statistics

[6] Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 1973; 33: 613619.

[7] Kraemer HC. Ramifications of a population-model for kappa as a coefficient of reliability. Psychometrika 1979; 44(4):461472.

[8] Commenges D, Jacqmin H. The intraclass correlation coefficient distribution-free definition and test. Biometrics 1994; 50:517526.

[9] Suen HK. Principles of Test Theories. LEA Publishers, Hilsdale, New Jersey, 1990.

[10] Spearman C. Correlation calculated from faulty data. British Journal of Psychology 1910; 3:271296.

[11] Brown W. Some experimental results in the correlation of mental abilities. British Journal of Psychology 1910; 3:296322.

[12] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16:297334.

[13] Kuder G, Richardson M. The theory of estimation of test reliability. Psychometrika 1937; 2:151160.

[14] Novick MR, Lewis C. Coefficient alpha and the reliability of composite measurement. Psychometrika 1967; 32:113.

[15] Raykov T. Scale reliability, Cronbach's coefficient Alpha, and violations of essential tau-equivalence with fixed congeneric components. Multivariate Behavioral Research 1997; 32(4):329353.

[16] Neter J, Wasserman W, Kutner MH. Applied Linear Statistical Models. Richard D. Irwin, Inc., Homewood, IL, USA, 1985

[17] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. The Danish Institute of Educational Research, 1960.

[18] van der Linden WJ, Hambleton RK (editors). Handbook of Item Response Theory. Springer-Verlag, New York, 1997.

[19] Samejima F. Estimation of Reliability Coefficients Using the Test Information Function and Its Modifications. Applied Psychological Measurement 1994; 18(3):229244.

[20] Andrich D. An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. Educational Research and Perspectives 1982; 9(1):95104.

[21] Martinková P. Reliability of Measurements Consisting of Dichotomously Scored Items. Unpublished dissertation. Charles University, Prague, 2007.

[22] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0, URL http://www.R-project.org.

[23] Cios KJ, Kurgan LA. SPECT heart data. Colorado, USA, 2010. URL http://archive.ics.uci.edu/ml/datasets/SPECT+Heart.

[24] Kurgan LA et al. Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial Intelligence in Medicine 2001; 23: 149-169.

Contact
RNDr. Patrícia Martinková, Ph.D.
Institute of Computer Science AS CR
Centre of Biomedical Informatics
Pod Vodárenskou věží 2
182 07 Prague 8
Czech Republic
e-mail: martinkova@euromise.cz