Reliability of Educational Tests

Martinková, Patrícia
2006

Reliability of Educational Tests

Post-Graduate Student:
MGR. PATRÍCIA MARTINKOVÁ
Department of Medical Informatics
Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2
182 07 Praha 8
Czech Republic
martinkova@euromise.cz

Supervisor:
DOC. RNDR. KAREL ZVÁRA, CSC.
Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University in Prague
Sokolovská 83
186 75 Praha 8
Czech Republic
karel.zvara@mff.cuni.cz

Field of Study:
Probability and Mathematical Statistics

The work was supported by the grant 1M06014 of the Ministry of Education of the Czech Republic.
It contains results published in [1], some recent results were added.

Abstract

The paper deals with reliability of measurements in the context of multiple-item testing instruments, such as educational tests. We concentrate on popular characteristic widely used for estimation of reliability called *Cronbach's alpha*, which is suited for normally distributed error term. Further we discuss modifications of Cronbach's alpha for the case of dichotomous (true-false) scoring.

1. Reliability

When describing the reliability of measurement, it is usually assumed that the measurement Y is composed out of two random variables: an unobservable true value T and an error term e,

$$Y = T + e.$$

The error term is supposed to have a zero mean $E(e) = 0$, a positive variance, and to be independent from the true value T. Therefore:

$$\text{var}\,(Y) = \text{var}\,(T) + \text{var}\,(e).$$

The *reliability* of such measurement is defined by:

$$R = \frac{\text{var}\,(T)}{\text{var}\,(Y)} = 1 - \frac{\text{var}\,(e)}{\text{var}\,(Y)} \tag{1}$$

and it compares variability of the error term with the variability of measured property. The smaller the error variance relative to the observed score variance, the more reliable is the measurement. Thus, the measurement is considered to be reliable when the value of reliability is close to 1.

Here, we should point out that reliability is sample-dependent. Therefore a certain test can have a different reliability when given to a population with a high variability of tested knowledge than when given to a population with a low variability of the knowledge.

The following simple lemmas give us a natural interpretation of the reliability.

Lemma 1.1 *Having two independent measurements $Y_1 = T + e_1, Y_2 = T + e_2$ of the same property T, where var (e_1) = var (e_2), the reliability can be expressed as the correlation between these two measurements, $R =$ corr (Y_1, Y_2).*

Proof:

$$\text{corr}\,(T + e_1, T + e_2) = \frac{\text{cov}\,(T + e_1, T + e_2)}{\sqrt{\text{var}^2(Y)}} = \frac{\text{cov}\,(T, T) + 0}{\text{var}\,(Y)} = \frac{\text{var}\,(T)}{\text{var}\,(Y)} = R. \qquad \square$$

In terms of educational tests, the reliability reflects to what extent it gives the same result when taken repeatedly by the same person under the same conditions.

Lemma 1.2 *The reliability can be expressed as the squared value of the correlation between the observed score and the true score, corr*$^2(Y,T)$.

Proof:

$$\text{corr}^2(Y,T) = \frac{\text{cov}^2(T+e,T)}{\text{var}(Y)\text{var}(T)} = \frac{\text{var}^2(T)}{\text{var}(Y)\text{var}(T)} = \frac{\text{var}(T)}{\text{var}(Y)} = R.$$

□

Thus, the reliability of an educational test measures the strength of the relationship between the score reached by a student and his/her true knowledge.

Unfortunately, none of these representations is useful when estimating the reliability of educational tests because they cannot be directly estimated from the observed data. We cannot estimate the error variance var(e), the true score T, nor the knowledge of a student by the same test twice and independently. Therefore, when estimating the reliability of an educational test, we mostly take into account a fact that such a test is a composite measurement.

2. Reliability of composite measurement

We consider the problem of measuring the reliability of multiple-item testing instrument, such as in educational test. Consider a series of items $Y_j = T_j + e_j$, for $j = 1,\ldots,m$, where the error terms e_j are mutually independent and independent on the true scores T_k for $k = 1,\ldots,m$, having the same variance var$(e_j) = \sigma_e^2$, and mean E$\,e_j = 0$. The observed overall score of the m items is given by $Y = Y_1, + \cdots + Y_m$ and the unobservable overall true score is given by $T = T_1, + \cdots + T_m$. The reliability of such a composite measurement is defined by (1) and with regard to the above mentioned assumptions can further be expressed as:

$$R_m = \frac{\text{var}(T)}{\text{var}(Y)} = \frac{\text{var}(T)}{\text{var}(T) + \text{var}(\sum e_j)} = \frac{\text{var}(T)}{\text{var}(T) + m\sigma_e^2}. \tag{2}$$

The next lemma gives a relationship between reliability of a composite measurement and reliability of an item in one special case:

Lemma 2.1 *If for the items' true score the following holds simultaneously:*

$$\textit{var}(T_1) = \cdots = \textit{var}(T_m) = \sigma_T^2$$

$$\textit{corr}(T_j, T_k) = 1, \qquad j,k = 1,\ldots,m,$$

then all the reliabilities R_1 of the items are equal and the reliability of the whole test can be expressed in ***Spearman-Brown formula****:*

$$R_m = \frac{mR_1}{1+(m-1)R_1} \tag{3}$$

Proof:

$$\begin{aligned}
\text{var}\left(\sum_{j=1}^m T_j\right) &= \sum_{j=1}^m \text{var}(T_j) + \sum\sum_{j\neq k} \text{cov}(T_j,T_k) = \\
&= m\sigma_T^2 + m(m-1)\sigma_T^2 = m^2\sigma_T^2, \\
\text{var}\left(\sum_{j=1}^m Y_j\right) &= \text{var}\left(\sum_{j=1}^m T_j\right) + \text{var}\left(\sum_{j=1}^m e_j\right) = m^2\sigma_T^2 + m\sigma_e^2.
\end{aligned}$$

Therefore

$$\begin{aligned} R_m &= \frac{\text{var}\left(\sum_{j=1}^m T_j\right)}{\text{var}\left(\sum_{j=1}^m Y_j\right)} = \frac{m^2\sigma_T^2}{m^2\sigma_T^2 + m\sigma_e^2} = \frac{m\frac{\sigma_T^2}{\sigma_T^2+\sigma_e^2}}{1+(m-1)\frac{\sigma_T^2}{\sigma_T^2+\sigma_e^2}} \\ &= \frac{mR_1}{1+(m-1)R_1}. \end{aligned}$$

□

Related to this lemma is a fact, that reliability of an educational test is dependent on the number of its items. Therefore, by adding suitable items to the test, the reliability could approach as close to 1 as we would desire. When comparing reliabilities of two educational tests, which in principle can't have the same number of items, we should bear this property of reliability in mind.

3. Cronbach's Alpha

As a measure of reliability in classical test theory, Cronbach [2] proposed the coefficient alpha. This characteristic estimates the consistency between items in a test and it is defined as:

$$\alpha_{CR} = \frac{m}{m-1}\frac{\text{var}\,(Y) - \sum_j \text{var}\,(Y_j)}{\text{var}\,(Y)} = \frac{m}{m-1}\frac{\sum\sum_{j\neq k}\sigma_{jk}}{\sum\sum_{j,k}\sigma_{jk}}, \tag{4}$$

where σ_{jk} is the covariance of the pair (Y_j, Y_k). Novick and Lewis [3] has shown that Cronbach's alpha is always a lower bound of the reliability

$$\alpha_{CR} \leq R$$

and is equal to reliability only if the conditions of Lemma 2.1 are fulfilled.

A very pleasant property of Cronbach's alpha is the fact that this characteristic is easy to estimate from the data simply by using sample variances and sample covariances instead of their population counterparts in (4). This sample estimate can further be rewritten (for proof see [4]) in terms of the two-way ANOVA as:

$$\hat{\alpha}_{CR} = \frac{MS_T - MS_e}{MS_T} = 1 - \frac{1}{F_T}, \tag{5}$$

where MS_T and MS_E are the mean sums of squares and F_T is statistics widely used for testing the hypothesis $\text{var}\,(T) = 0$ when normality of variables can be assumed.

Notation (5) gives important properties of our estimate:

- $\hat{\alpha}$ can take values between $-\infty$ and 1, although only positive values make sense for reliability.
- The greater the estimate of reliability is, the better the educational test can distinguish between the students. This points out the fact, that Cronbach's alpha was designed as a coefficient of internal consistency.
- The estimate equals one, if and only if there exist constants $a_i, b_j, i = 1, \ldots, n, j = 1, \ldots, m$, so that the score reached by the i-th student in the j-th item can be written as $a_i + b_j$. This means that in this case, to get all the information about students, one item would be enough. Therefore, when getting too high an estimate of Cronbach's alpha, one should actually think of lowering the number of items.

4. Cronbach's alpha for dichotomous items

In fact, Cronbach's alpha was designed as a generalization of the so called Kuder-Richardson formula 20 for dichotomous scoring, already proposed in 1937 in [6]:

$$\hat{\alpha} = \frac{m}{m-1}\frac{s^2 - \sum_{j=1}^m p_j(1-p_j)}{s^2}, \tag{6}$$

where p_j is a relative frequency of correct answers to the jth item and s^2 is a sample estimate of the variance of total scores. One can easily see that (6) can be obtained when computing the sample estimate of Cronbach's alpha (4) in the case of dichotomous scoring, where $\hat{\mathrm{E}}\, Y_j = p_j$ is the proportion of correct answers to the jth question and $\hat{\mathrm{var}}\,(Y_j) = p_j(1-p_j)$.

Nevertheless, with dichotomous items, the assumptions of analysis of variance are violated. The scores cannot be assumed to have normal distribution, and moreover, the variance is dependent on the mean value. Therefore it is a matter of question to what extent is this estimate appropriate at all.

4.1. Proposed modifications of Cronbach's alpha estimate

Formula (5) led Zvara [5] to the idea of modifying Cronbach's alpha for the case of binary outcomes by replacing F_T by statistics used for testing the hypothesis $H_0 : \mathrm{var}\,(T) = 0$ in logistic regression. This is equal to testing the submodel B where the score Y_{ij} depends only on the test item (and doesn't depend on the student's ability) against the model A+B where the score Y_{ij} depends on the student and on the test item. Appropriate statistics is the difference of deviance in the submodel and in the model $X^2 = D(B) - D(A+B)$, which has under the null hypothesis the $\chi^2(n-1)$ distribution. Therefore, the proposed estimate is:

$$\hat{\alpha}_{log} = 1 - \frac{n-1}{X^2}. \tag{7}$$

In this work we are trying to justify the estimate (7), so far called the *logistic estimate of Cronbach's alpha* or shortly *logistic alpha*, and to demonstrate its qualities by simulations.

5. Extended beta-binomial model

The model used most often for describing items with dichotomous scoring is the logit-normal model called Rasch model [7], [8]. In this model the probability of a correct response of person i on item j is given by:

$$\mathrm{P}\,(Y_{ij} = y_{ij}; \pi_i, \delta_j) = \frac{\exp[y_{ij}(\pi_i + \delta_j)]}{1 + \exp(\pi_i + \delta_j)}, \tag{8}$$

where π_i describes the level of ability of person i and δ_j is an unknown parameter describing the difficulty of item j.

Evaluating the true reliability of a composite measurement of items which obey Rasch model with certain parameters $\pi_i, i = 1, \ldots, n$, and $\delta_j, j = 1, \ldots, m$ is a difficult task. That is why for simulations we propose the *extended beta-binomial model*, where calculating the true reliability is tractable (see (9)). The motivation of this model is following:

An often used model in reliability studies of binary data (see for example [9], [10]) is the *beta-binomial model*. In this model we assume, that the probability of success π_i varies over subjects $i = 1, \ldots, n$ according to a beta distribution with parameters a and b, and, conditional to this probability, the total score Y_i of the ith person is binomially distributed. Choice of beta distribution for π_i is logical since it is a flexible distribution and leads to mathematically tractable results. A pleasant property of the beta-binomial model is the fact, that the first two moments for the total score are easy to compute:

$$\mathrm{E}\,(Y) = n\mu = n\frac{a}{a+b}$$

$$\mathrm{var}\,(Y) = n\mu(1-\mu)\left[1 + (n-1)\frac{\theta}{1+\theta}\right],$$

where μ is the marginal probability of success for any individual, $\theta = \frac{1}{a+b}$ and $\frac{\theta}{1+\theta} = \rho$ is the intraclass correlation $\mathrm{corr}\,(Y_{ij}, Y_{ik})(j \neq l)$ common for any subject and any pair of responses.

An unpleasant property of this model for our situation is the fact that it does not allow for different difficulties of items. Hand in hand with this goes the common-correlation structure which is impossible in our case.

When trying to extend for different difficulties of items and yet preserve the structure of the beta-binomial model, we can think of the following model: We assume again, that the probability of success π_i varies over subjects $i = 1, \ldots, n$ according to a beta distribution with parameters a and b. We qualify the impact of the difficulty of the jth item by a small number δ_j, assuming that $\sum_{j=1}^{m} \delta_j = 0$. When parameters a, b are large enough, there is a slight danger that the sums $\pi_i + \delta_j$ get outside the interval $(0, 1)$. Therefore, $Y_{i1}, \ldots, Y_{im}$ are for a given π_i independent random variables with alternative distribution alt$(\pi_i + \delta_j)$. The total scores Y_i are sums of such random variables.

5.1. Properties of Y_{ij} in the extended beta-binomial model

For conditional mean and variance, it holds

$$\begin{aligned}
\mathrm{E}\,(Y_{ij}|\pi_i) &= \mathrm{E}\,(Y_{ij}^2|\pi_i) = \mathrm{P}\,(Y_{ij} = 1|\pi_i) = \pi_i + \delta_j, \\
\mathrm{var}\,(Y_{ij}|\pi_i) &= \mathrm{E}\,(Y_{ij}^2|\pi_i) - (\mathrm{E}\,(Y_{ij}|\pi_i))^2 = (\pi_i + \delta_j)(1 - (\pi_i + \delta_j)).
\end{aligned}$$

Therefore the unconditional mean is

$$\mathrm{E}\,(Y_{ij}) = \mathrm{E}\,\mathrm{E}\,(Y_{ij}|\pi_i) = \frac{a}{a+b} + \delta_j = \mu + \delta_j,$$

where we assigned $\mu = a/(a+b)$ for the mean value of the beta distribution. For the unconditional variance it holds:

$$\begin{aligned}
\mathrm{var}\,(Y_{ij}) &= \mathrm{var}\,(\mathrm{E}\,(Y_{ij}|\pi_i)) + \mathrm{E}\,(\mathrm{var}\,(Y_{ij}|\pi_i)) \\
&= \mathrm{var}\,(\pi_i + \delta_j) + \mathrm{E}\,((\pi_i + \delta_j)(1 - (\pi_i + \delta_j))) \\
&= \frac{ab}{(a+b)^2(a+b+1)} + \int_0^1 (\pi + \delta_j)(1 - (\pi + \delta_j)) \frac{1}{\mathrm{B}(a,b)} \pi^{a-1}(1-\pi)^{b-1} \mathrm{d}\pi \\
&= \frac{ab}{(a+b)^2(a+b+1)} + (\mu + \delta_j) - \frac{a(a+1)}{(a+b)(a+b+1)} - 2\delta_j \frac{a}{a+b} - \delta_j^2 \\
&= \mu(1-\mu) + \delta_j(1 - 2\mu - \delta_j).
\end{aligned}$$

Because $\rho = \frac{\theta}{1+\theta} = \frac{1}{a+b+1}$, the covariance of variables Y_{ij}, Y_{it} for $j \neq t$ equals

$$\begin{aligned}
\mathrm{cov}\,(Y_{ij}, Y_{it}) &= \mathrm{cov}\,(\mathrm{E}\,(Y_{ij}|\pi_i), \mathrm{E}\,(Y_{it}|\pi_i)) \\
&= \mathrm{cov}\,(\pi_i + \delta_j, \pi_i + \delta_t) = \mathrm{var}\,(\pi_i) \\
&= \frac{ab}{(a+b)^2(a+b+1)} = \rho\mu(1-\mu),
\end{aligned}$$

Let's define

$$C_j = 1 + \delta_j \frac{1 - 2\mu - \delta_j}{\mu(1-\mu)}.$$

Then the correlation between Y_{ij} and Y_{it} for $j \neq t$ is

$$\mathrm{corr}\,(Y_{ij}, Y_{it}) = \frac{\mathrm{cov}\,(Y_{ij}, Y_{it})}{\sqrt{\mathrm{var}\,Y_{ij}\,\mathrm{var}\,Y_{it}}} = \rho \frac{1}{\sqrt{C_j C_t}}.$$

For constant difficulties of items $\delta_j = 0$ we get the common correlation structure, corr $(Y_{ij}, Y_{it}) = \rho$. For unequal difficulties of items it is natural to assume $a = b$ (to assume symmetric distribution of knowledge), therefore $\mu = 1/2$. In this case $C_j = 1 - 4\delta_j^2$, thus the impact of $\delta < 1$ is small.

5.2. Properties of total scores Y_i in the extended beta-binomial model

The total score of the ith student is the total number of correctly answered items $Y_i = \sum_{j=1}^{m} Y_{ij}$. We get

$$\begin{aligned}
\mathrm{E}\,(Y_i|\pi_i) &= m\frac{1}{m}\sum_{j=1}^{m}(\pi_i+\delta_j) = m\pi_i,\\
\mathrm{var}\,(Y_i|\pi_i) &= \mathrm{var}\,(\sum_{j=1}^{m} Y_{ij}|\pi_i) = \sum_{j=1}^{m}(\pi_i+\delta_j)(1-(\pi_i+\delta_j) = m\frac{1}{m}\sum_{j=1}^{m}(\pi_i+\delta_j) - m\frac{1}{m}\sum_{j=1}^{m}(\pi_i+\delta_j)^2\\
&= m\left(\pi_i - \left(\frac{1}{m}\sum_{j=1}^{m}(\pi_i+\delta_j)\right)^2\right) - m\left(\frac{1}{m}\sum_{j=1}^{m}(\pi_i+\delta_j)^2 - \left(\frac{1}{m}\sum_{j=1}^{m}(\pi_i+\delta_j)\right)^2\right)\\
&= m\pi_i(1-\pi_i) - m\kappa_\delta,
\end{aligned}$$

where $\kappa_\delta = \frac{1}{m}\sum_{j=1}^{m}\delta_j^2$. Therefore it holds:

$$\begin{aligned}
\mathrm{E}\,(Y_i) &= \mathrm{E}\,\mathrm{E}\,(Y_i|\pi_i) = m\mathrm{E}\,(\pi_i) = m\frac{a}{a+b} = m\mu,\\
\mathrm{var}\,(Y_i) &= \mathrm{var}\,(m\pi_i) + \mathrm{E}\,(m\pi_i(1-\pi_i) - m\kappa_\delta) = m^2\mathrm{var}\,(\pi_i) + m\mathrm{E}\,(\pi_i) - m\mathrm{E}\,(\pi_i^2) - m\kappa_\delta)\\
&= m^2\frac{ab}{(a+b)^2(a+b+1)} + m\frac{a}{a+b} - m\frac{a(a+1)}{(a+b)(a+b+1)} - m\kappa_\delta\\
&= m\mu(1-\mu)(1+(m-1)\rho) - m\kappa_\delta.
\end{aligned}$$

Finally, we are getting to the **reliability of the total score** Y_j in the extended binomial model. We define it according to [11] as a fraction of variability between students (variability of conditional mean values $\mathrm{E}\,(Y_i|\pi_i)$) and variability of students' total scores Y_i:

$$\begin{aligned}
R_m &= \frac{\mathrm{var}\,(\mathrm{E}\,(Y_i|\pi_i))}{\mathrm{var}\,(Y_i)} = \frac{\mathrm{var}\,(m\pi_i)}{\mathrm{var}\,(Y_i)} = \frac{m^2\mu(1-\mu)\rho}{m\mu(1-\mu)(1+(m-1)\rho) - m\kappa_\delta}\\
&= \frac{m\rho}{1+(m-1)\rho - \frac{\kappa_\delta}{\mu(1-\mu)}}.
\end{aligned} \tag{9}$$

When the difficulties of items are all equal $\delta_j = 0$, we get the well known Spearman-Brown formula (3). For unequal difficulties of items, the reliability of total scores is a bit larger.

Formula (9) is very important for simulations. For the given parameters of beta-binomial distribution a, b, and for the given difficulties δ_j, $j = 1, \ldots, m$ we can calculate the true reliability R_m and compare it with estimates calculated from simulated data.

6. Simulations

So far, a single simulation was done. We investigated the behavior of the classical and logistic Cronbach's alpha estimator in the extended beta-binomial distribution via simulation for number of items $k = 11$, number of students $n = 20$ and items' difficulties δ_j equidistantly distributed between -0.1 and 0.1. The parameters $a = b$ of the beta-binomial distribution were chosen from the interval $\langle 1, 15\rangle$ with step 0.2. For each simulation we generated 100 data sets and computed bias and mean squared error of the classical estimates of Cronbach's alpha and of the logistic estimates of Cronbach's alpha. For each out of 71 possible values of parameters $a = b$, also the theoretical value of reliability was evaluated, using the equation (9).

In Figure 1, the bias and mean squared error of classical and logistic estimate of Cronbach's alpha is shown for different values of the true reliability.

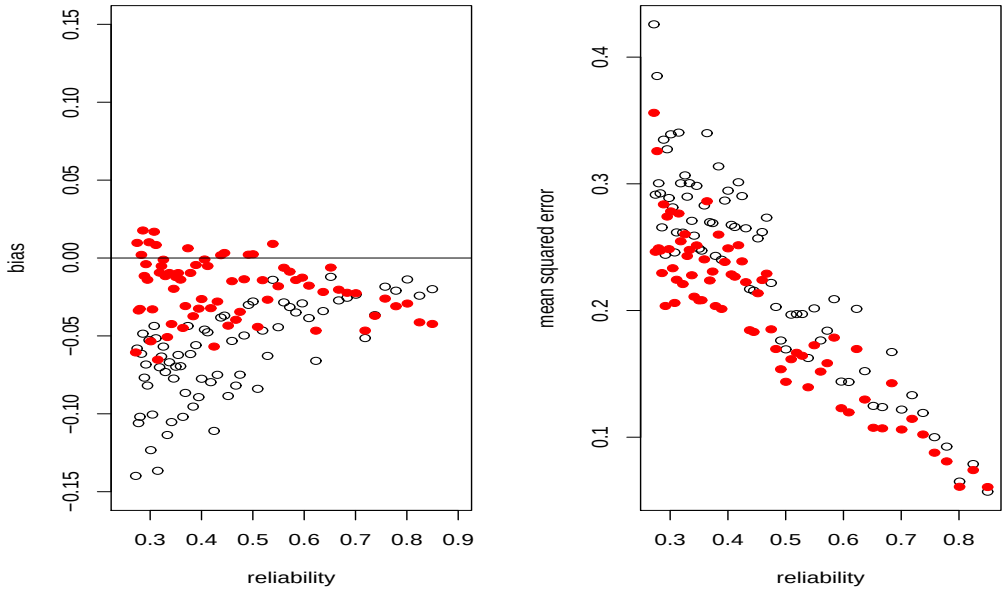

Figure 1: The bias and mean squared error of classical (circle) and proposed logistic (solid circle) estimate of α_{CR}

According to Figure 1, the proposed logistic alpha performs as an estimate of reliability better than the classical Cronbach's alpha estimate in the extended beta-binomial model. The logistic alpha tends to give worse results only for high true reliabilities, thus for small a, b, which is the case of high probability of cutting in the extended beta-binomial model.

7. Conclusions and Discussion

According to our simulations, the proposed logistic estimate of Cronbach's alpha performs better for binary data of the extended beta-binomial model than the classical Cronbach's alpha estimate.

When going through section 5, one can conclude, that there is no need for beta distribution in the extended model to get the same formula for true reliability of total scores Y_i. Therefore, more complex simulations in this class of models should be done.

Also, the remaining task is to justify the proposed class of models for the real data (would Hosmer-Lemeshow goodness-of-fit test work?). Or even better to justify the proposed estimate (7) for the Rasch model (8).

For testing the hypothesis $H_0 : \mathrm{var}\,(T) = 0$, there also exist other statistics besides difference of deviances. Therefore, other modifications of classical Cronbach's alpha estimate could also be defined and compared with logistic alpha discussed in this article.

References

[1] P. Martinková, K. Zvára jr., J. Zvárová, K. Zvára "The New Features of the ExaMe Evaluation System and Reliability of Its Fixed Tests", *Methods of Information in Medicine*, vol. 45, pp. 310–315, 2006.

[2] L. J. Cronbach "Coefficient Alpha and the internal structure of tests", *Psychometrika*, vol. 16, pp. 297–334, 1951.

[3] M. R. Novick, Ch. Lewis *Coefficient Alpha and the Reliability of Composite Measurements*, Educational Testing Service, Princeton, New Jersey, 1966.

[4] P. Rexová, *Spolehlivost měření* [Reliability of measurements, In Czech] Diploma thesis. Department of Probability and Mathematical Statistics, Charles University, Prague 2003.

[5] K. Zvára "Měření reliability aneb bacha na Cronbacha". *Statistický bulletin*, vol. 13, pp. 13–20, 2002.

[6] G. Kuder, M. Richardson "The theory of estimation of test reliability". *Psychometrika*, vol. 2, pp. 151-160, 1937.

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

[8] P. Rexová "Item Analysis of Educational Tests in System ExaMe". Doktorandský den '04. ICS AS CR, 2004.

[9] M. S. Ridout, C. G. B. Demeterio, D. Firth "Estimating intraclass correlation for binary data". *Biometrics*, vol. 55, pp. 137–148, 1999.

[10] G. Zou, A. Donner "Confidence Interval Estimation of the Intraclass Correlation Coefficient for Binary Outcome Data". *Biometrics*, vol. 60, pp. 807–811, 2004.

[11] D. Commenges, H. Jacqmin "The Intraclass Correlation Coefficient Distribution-Free Definition and Test". *Biometrics*, vol. 50, pp. 517–526, 1994.