ON THE RELIABILITY OF DIDACTIC TESTS

Violeta Marinova, Lybomir Christov, Dimiter Tsvetkov

CLASSICAL TEST THEORY (CTT)

Usually a scale consists of a fixed number of items of the following type: binary scale with "yes/no" answers, multiple choice scales when examinee should choose the right answer between some alternatives and graded response scale. In all cases the raw score is the sum of the item scores. It is commonly accepted that the individual of higher raw score owns the measured construct in a higher rate.

Each scale should be reliable and valid. The reliability means that the scale measures relatively exact some construct. In the low reliable scales the raw score is more a result of a system error instead of meaningful measurement. The validity means that the scale measures just the construct for which the scale is assembled. Low valid scales are in fact useless. Note that the validity can not exceed its reliability.

The main issues of the classical test theory are related with the ideas of reliability and validity. The basic advantage of CTT is that its assumptions are almost trivial nevertheless that its main assumption – existence of parallel scales (parallel tests) brings a hypothetical nature. These characteristics of CTT along with its transparency guarantee its importance instead of further generalizations.

The CTT assumes that the observed score X_m and the true score T_m obey the following simple relation

$$X_m = T_m + E_m ,$$

where E_m is the error score – the result of some random influence. Here the index m denotes the measurement associated with a given scale. The observed score is what we get during the measurement and the true score is exact what we want to measure. The basic hypothesis of CTT are: ***the expectation of the error is zero,*** $\mathbf{E}[E_m] = 0$, ***the linear correlation between the error and the true score is zero*** $\mathbf{r}[E_m, T_m] = 0$, ***the linear correlation between the error of a given measurement and the true score of another one is zero*** $\mathbf{r}[E_m, T_{m'}] = 0$, ***the linear correlation between the error of a given measurement and another is zero,*** $\mathbf{r}[E_m, E_{m'}] = 0$.

It is easy to see that $\mathbf{E}[X] = \mathbf{E}[T]$.

The distinct measurements m and m' are called ***equivalent*** when they yield the same result on the persons and the distribution of the errors are equal and independent. The equivalent measures have identical probability characteristics.

The distinct measurements m and m' are called ***parallel*** when they yield equal results on the persons and the dispersions of the errors are equal – $\sigma[E_{mp}] = \sigma[E_{m'p}]$. If we reject the latter assumption we get the so-called τ-***equivalent*** measurements. The τ-equivalent measurements m and m' differ only in their error distributions. Note that $\mathbf{D}[X] = \mathbf{D}[T] + \mathbf{D}[E]$.

The value

$$\mathbf{rel}_X = (\mathbf{r}_{XT})^2 = \mathbf{r}_{XX'}$$

where X and X' are parallel measurements is called ***reliability*** of the measurement (test) and the value is the ***reliability index***.

The ***validity*** coefficient of the measurement with respect to the measurement is called the value

$$\mathbf{val}[X|Y] = \mathbf{r}_{XY} = \frac{\sigma[X,Y]}{\sigma[X]\sigma[Y]}$$

In fact the reliability is exact the validity of a given measurement with respect to itself or with respect to some other parallel measurement – $\mathbf{rel}[X] = \mathbf{r}_{XX'} = \mathbf{val}[X|X']$ where X and X' are parallel. In any case the validity is defined only with respect to some other measurement.

Let X be some measurement. Then for the linear regression line of the true score on the observed score we have $t = \alpha x + \beta$, where for the coefficient α and β it can be shown that $\alpha = \mathbf{rel}[X]$ and $\beta = (1 - \mathbf{rel}[X])\mu_X$. In this way for the regression line we have

$T = \mathbf{rel}_X X + (1 - \mathbf{rel}_X)\mu_X$ **.**

The latter is a strong argument for the main importance of the reliability. If the reliability is closer to zero then the observed score do not bring any information about the true score. Note that the regression line of the observed score on the true score is simply $X = T$**.**

Let $X = T_X + E_X$ and $Y = T_Y + E_Y$ are distinct measurements. Then it can be shown that $\mathbf{r}_{XY} \le \mathbf{r}_{XT_X} = \sqrt{\mathbf{r}_{XX'}}$, i.e. the linear correlation between observed scores do not exceed the reliability index.

Consider parallel measurements Y_1, Y_2, Y_1', Y_2' and the composite scale $X = Y_1 + Y_2$ and $X' = Y_1' + Y_2'$. Then it holds

$$\mathbf{r}(X, X') = \frac{2\mathbf{r}(Y, Y')}{1 + \mathbf{r}(Y, Y')},$$

that yields the so-called ***Spierman-Brown*** formula

$$\mathbf{r}_{XX'} = \frac{2\mathbf{r}_{YY'}}{1 + \mathbf{r}_{YY'}} .$$

Here $\mathbf{r}_{XX'}$ is the reliability of X and $\mathbf{r}_{YY'}$ is the reliability of the Y compounds.

The Spierman-Brown formula has a main importance in the applications. The given scale is separated (in an arbitrary way) in two scales with approximately equal number of items. In this case the overall scale reliability is estimated in the following way

$$\mathbf{rel} = \frac{2\mathbf{r}}{1+\mathbf{r}},$$

where $\mathbf{r}$ is the correlation coefficient between the subscales.

Now consider some distinct measurements Y_i, $i = 1,2,\ldots,n$, and let X be the sum composite measurement, i.e. $X = Y_1 + Y_2 + \cdots + Y_n$. Then it can be shown that

$$\mathbf{rel}_X \geq \alpha,$$

where the value α, defined by the formula

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sigma_{Y_1}^2 + \sigma_{Y_2}^2 + \cdots + \sigma_{Y_n}^2}{\sigma_X^2}\right],$$

is the so-called ***Cronbachs alpha coefficient***. The equality $\mathbf{rel}_X = \alpha$ is attained only in the case of essentially τ-equivalent Y_1, Y_2, ..., Y_n, i.e. when $T_i = T_j + a_{ij}$ for some constants a_{ij}.

This coefficient is very easy to calculate and is the most commonly used in the practice to estimate the reliability.

When $Y_1, Y_2, ..., Y_n$ are parallel to Y we have the following generalization of the Spierman-Brown formula

$$\mathbf{r}_{XX'} = \frac{n\mathbf{r}_{YY'}}{1+(n-1)\mathbf{r}_{YY'}},$$

that shows how the reliability increases by adding more items in the scale.

If the scale consists of n binary items with observed frequencies of the correct answers p_i, $1 \leq i \leq n$, then $\sigma^2(Y_i) = p_i q_i$ where $q_i = 1 - p_i$ and the formula for the Cronbach alpha reduces to

$$\alpha_{(20)} = \frac{n}{n-1}\left[1 - \frac{p_1 q_1 + p_2 q_2 + \cdots + p_n q_n}{\sigma_X^2}\right]$$

which is called formula-20 of ***Kuder-Richardson***.

In the same way we have that $\mathbf{rel}_X \geq \lambda_2$ where λ_2 is the ***Guttman reliability index***

$$\lambda_2 = 1 - \frac{\sum_{i=1}^{n}\sigma^2(Y_i)}{\sigma_X^2} + \frac{\sqrt{\frac{n}{n-1}\sum_{i\neq j}\sigma^2(Y_i, Y_j)}}{\sigma_X^2}.$$

The point estimate of the alpha coefficient is given by the formula

$$\hat{\alpha} = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} s_i^2}{s_X^2}\right],$$

using the given data after the measurement.

ITEM ANALYSIS

The composite scales consist of items. Obviously the quality of a given scale depends on the characteristics of the compound items and their interrelations. Let X be a composite scale of items Y_1, Y_2, ..., Y_n, i.e. $X = Y_1 + Y_2 + \cdots + Y_n$, and let we have a sample of N persons (examinees). Then for the average score we have $\bar{x} = p_1 + p_2 + \cdots + p_n$, where

$$p_i = \frac{1}{N}\sum_j y_{ij}$$

is the average score of the examinees on the i-th item. The value p_i is known as the ***item difficulty*** and obviously is a statistical estimate for the theoretical mean π_i. For example in the case of binary items, is the proportion of the correct answers. It is clear that the item with very high or very low difficulty is not enough consistent with the scale. The good difficulty is between 0.3 and 0.7 or in the worse case between 0.1 and 0.9.

The discriminate rate of a given item is defined as the ability of the item to discriminate between the good and bad results. In this sense as a discrimination index is used the correlation coefficient $\mathbf{r}_{XY_i}$ between the item score and the overall scale score. For the alpha coefficient we have

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_X \sum_{i=1}^{n}\sigma_i \mathbf{r}_{Y_i X}}\right] = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sum_{i=1}^{n}\sigma_i^2 + \sum_{i\neq j}\sigma_i\sigma_j \mathbf{r}_{Y_i Y_j}}\right],$$

which shows that the higher discrimination indexes yield a higher values of the α coefficient. By this reason we should choose items with higher $\mathbf{r}_{XY_i}$ values. The formula above shows also that the items should be of relatively high mutual correlations. On the other hand for the validity index of X with respect to some criterion Z which is the correlation $\mathbf{r}_{XZ}$ of X and Z we have

$$\mathbf{r}_{XZ} = \frac{\sum_{i=1}^{n}\sigma_i \mathbf{r}_{Y_i Z}}{\sqrt{\sum_{i=1}^{n}\sigma_i^2 + \sum_{i\neq j}\sigma_i\sigma_j \mathbf{r}_{Y_i Y_j}}}$$

The last formula exhibits that to receive higher validity we must choose items with lower mutual correlations which as we already pointed out leads to lower reliability.

This conclusion teaches us that the "ideal" items are with high discrimination coefficient and with low mutual correlations. ***This requirement is not easy to achieve and may be represents itself the art of the test construction.***

ITEM RESPONSE THEORY (IRT)

In the IRT models the probability for correct answer depends on the proficiency level of the examinee and on some additional item parameters. For free-answer items the Rasch model is often used. In this model the probability of a correct answer to the -th item is due to the formula

$$P_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}},$$

where b_i the difficulty parameter. The Rasch model assumes equal item discrimination and no probability for the answer guessing. These assumptions are very strong for the case of multiple choice items. On the other hand the Rasch model is very stable with respect to the various violations (and gives certain results even in the case of meaningless scales). The main assumption here and everywhere is the assumption of one-dimensional latent space, i.e. that the scale is aimed to measure (and in fact measures) the only one construct.

For the multiple choice items the most proper model is the Birnbaum's model in which the probability of a correct answer to the *i*-th item is due to the formula

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}},$$

where b_i is the difficulty parameter, a_i stand for the discrimination parameter and c_i represent the guessing level probability.

These models should be calibrated under a given data from the test performance. The items parameter estimation is usually done by maximal likelihood principle or by marginal maximal likelihood principle. The estimation is possible only by using of some computer programs which can be found for example in the site www.assess.com. The authors of the present paper have their own software that can estimate the coefficients for the models mentioned (and also for other IRT models).

It is commonly accepted that the good calibration of a certain IRT model first of all shows that the model is reliable. A good calibration means that the numerical process converges well and does not show unrealistic parameter values.

Finally remember that a scale (test) which is not reliable enough is useless in the practice and the main purpose of this paper is to pay attention to the home-test-manufacturers on the last highly important fact.

REFERENCES

1. *Lord, F. M., Novik, M. R.* Statistical Theories of Mental Test Scores, 1968.

2. *Lord, F. M.* Applications of Item Response Theory to Practical Testing Problems, 1980.

3. *Hambleton, R., and H. Swaminathan*. Item Response Theory: Principles and Appli-cations, 1984.

НАДЕЖНОСТ НА ДИДАКТИЧЕСКИТЕ ТЕСТОВЕ

ВИОЛЕТА МАРИНОВА, ЛЮБОМИР ХРИСТОВ, ДИМИТЪР ЦВЕТКОВ

Резюме

Дидактическите тестове можгат да бъдат оценявани за надеждност и валидност. Надеждността отчита вътрешни закономерности на измерваните характеристики, което определя и нейната значимост. От друга страна надеждността има отношение към корелацията между наблюдаваните величини и техните истински стойности. Тестовете с ниска надеждност са практически безполезни. В класическата теория на тестовете оценки се правят предимно с помощта на коефициента алфа на Кронбах и коефициента на Гутман. В съвременната теория на тестовете се разглеждат някои вероятностни модели с цел постигане на по-висока надеждност. Като най-често използвани са 3-параметричния модел на Бирнбаум и в частност 1-параметричния модел на Раш.

В настоящата статия дискутираме някои практически правила, които могат да бъдат в помощ на тестващия за оценка надеждността на провеждани от него тестове. Те могат да бъдат използвани също така и при стандартизиране на тестове, приложими както към различни популации така и при наличие на различи други условия свързани с провеждане на тестовете.

ON THE RELIABILITY OF DIDACTIC TESTS

VIOLETA MARINOVA, LYBOMIR CHRISTOV, DIMITER TSVETKOV

Summary

Didactic tests should be reliable and valid. The reliability is the self-validity of a given measurement tool therefore there is no doubt that it is of a main importance in any didactic measurement. On the other hand the reliability represents the correlation between the test "observed score" and the test "true score" of the construct under measurement. Given a test with a poor reliability is useless. One can investigate reliability by means of various classical reliability indexes like the Kronbach alpha or the Guttman split-half reliability coefficient. Modern test theory (Item Response Theory) offers some probability models which purpose in fact is to achieve the test reliability to some extent. The most common used model is the 3-parametric Birnbaum model in the case of a test with multiple-choice item type. The most simple 1-parametric model (the Rasch model) is also used in the case of a test with free-answer items type.

Here we discuss some practical rules by help of which the teacher can verify the reliability of a given test. This question arises always when the teacher manufactures some achievement or criterion test for its own use. Another situation, when the reliability is of main importance, is the case when the teacher want to use given test which is standardized with respect to different conditions or the test is designed for different population.