ARGUMENTA OECONOMICA
No 2 (27) 2011
PL ISSN 1233-5835

Piotr Tarka *

MEASUREMENT, RELIABILITY AND SCALES CONSTRUCTION IN A VIEW OF CLASSICAL TRUE-SCORE THEORY

The main objective of this article was to describe issues related to theory of measurement and discuss significant role of selected reliability methods in a view of Classical True-Score Theory – CTM. In consequence the emphasis was put on description and comparison between different methods of reliability assessment such as: test-rest, parallel-test, split-half or internal consistency. The author investigated them primarily in context of psychometrics and its general applications in the area of marketing and customers research studies. In the second part of article, a new perspective on the measurement (Item Response Theory – IRT) was discussed. Later on, both concepts of measurement CTM and IRT were compared. Although researchers in the field of psychometrics have paid a considerable attention to measurement theory, in another field (e.g. marketing) these topics somehow have been neglected. And the recent advances in statistical analysis have drawn increasing attention to these nagging problems of measurement. Therefore in the article the author decided to review and stress the great importance of various concepts in reliability measurement. The author hopes that for researchers in the above field who wish to familiarize themselves with current debates over the right choice of an appropriate measurement design and strategies, it will be a good starting point for their own research and reliability estimation, especially when making a decision on how to develop an appropriate scale for measurement and choose for that scale respective estimate reliability. As a result, the description will be useful for managers, marketers, who want to study reliability and problems associated with scales construction to study customers behavior or market trends and those who want to attempt integrate measurement theory into their aggregate models of the business.

Keywords: reliability, methods of reliability, classical true-score theory, item response theory, scales

INTRODUCTION

The notion that measurement is crucial to science seems to be a commonplace and unexceptional observation. The most popular **definition of measurement** (as far as the social sciences measurement area is concerned) – is provided by Stevens (1951) explaining that *"measurement is the assignment of numbers to objects or events according to specific rules"*.

* Department of Marketing Research, Poznań University of Economics

This concept of measurement unfortunately lacked appropriate accuracy level. Duncan commented measurement (as opposed to Steven's description) more precisely adding *"measurement is also the assignment of numerals in such a way as to correspond to different degrees of a quality or property of some object or events"* (Duncan, 1984). More broadly stated – it assigns a numerical scale according to the size, value, or other characteristic of a tangible or intangible object. In effect and in practice of research one has obtained the scale which could be ranging as far as from: 0 to 1 (bad or good), 0 to 10 (as in athletic competitions), 1 to 5 (as in Likert scale being a part of psychometrics area measurement).

Moreover in Duncan's point of view, *"all measurement is the social measurement"*. This could be a reference to the earliest formal social measurement processes such as voting or census-taking. He further notes that *"their origins seem to represent early attempts to meet every day human needs, not merely experiments undertaken to satisfy scientific curiosity"*. He continues, saying that similar processes can be drawn in the history of physics where measurement of length or distance, area, volume, weight and time was conducted by ancient people in the course of solving practical and social problems. Physical science was built on the foundations of those achievements (Duncan, 1984).

In social sciences measurement and scales development (associated for example with marketing – in particular customer traits), problems with the measurement (according to Steven's expression) is that many of the customers phenomenon to be measured (e.g. analysis of their attitudes or preferences to products or services) are intangible and quite often are too abstract in order to be adequately characterized as either simple objects. In short they cannot be seen or touched. However, having based on Blalock (1968), and Zeller et al. (1979) and other authors findings, one could have at least attempted to create a general approach for their measurement and then scale development.

1. CONSTRUCTS AND DIFFERENT APPROACHES TO MEASUREMENT

The theory of construct plays a significant role on how we conceptualize our measure problems in science. In this context, it is worth mentioning that scientists tend to rely on numerous theoretical models that concern rather narrowly circumscribed scientific phenomenon. Very often measuring

elusive, intangible phenomenon (such as customer values, needs or personal traits) is derived from multiple, quickly evolving theories. As a result everything depends on: 1) the researcher's knowledge of optional theoretical construct, 2) choice of different (available) assessment strategies and 3) type of recognized trait whether **directly observable** (e.g. length of the board, height or weight) or **indirectly observable** (e.g. human beliefs, motivational states, value systems, expectancies, needs, emotions, or perceptions). Most indirect measures are sometimes called **indicators.**

Observable traits which are reflected by latent variables (whose presence is for example inferred from the pattern of covariation among the indicators) are called **reflective indicators.** They measure latent variable along with some measurement error. Usually measurement is undertaken on the basis of *summated scale*, where items represent reflective observable variables. Some part of these items may be eliminated from the scale without greater damage for the considered latent variable (e.g. a designed scale).

Bollen (1989) noted that, in some instances, indicators of a latent variable are uncorrelated and, rather than reflecting the variable, they cause it. Therefore if observable variables affect latent variables, then they are termed as **formative indicators.** Formative indicators are measured already without errors, and they are formed objectively, because their numbers do not depend completely on evaluation given by person. This type of measurement instruments is sometimes interchangeably called **index**. In practice, formative indicators are somewhat rare, perhaps because many constructs are not typically conceptualized with reference to their causes. In traditional customers research or psychometric criteria, a common practice is dominated by reflective indicators.

2. CONSTRUCTS DIMENSIONALITY AND SCALES DEVELOPMENT

For some reasons one can differentiate constructs according to their dimensionality. They can be varied either in **unidimensional** or **multidimensional** form. In customer research, many constructs such as human beliefs, motivational states, value systems, expectancies, needs, emotions, or perceptions are far and wide more complex and most of them belong to multidimensional description. If this is the case, part of the researcher's task is to decide on how finely these constructs can be extracted and described (Spector, 1992).

Constructs that are fairly homogenous are typically **unidimensional,** and constructs which have broad aspects or dimensions will be **multidimensional.** The development of **multidimensional scales** are not much different from their unidimensional counterparts. Procedures of the multidimensional scales development are the same as for unidimensional ones. On the conceptual end, the various components are specified. These components do not have to be unrelated, and often subscales of multidimensional instruments inter-correlate. However, conceptually they should be distinct.

The term "scale" reflects primarily a multi-item scales. In the process (sometimes referred to as “test”) of multi-item scale construction we usually have N persons taking a test that consists of k items where a score x_{ji} is given to the j-th person on the i-th item. The process of scale development containing multi-items typically involves the following stages (Malhotra, 2009) where:

1) scale cannot be developed until it is clear exactly what that scale is intended to measure;

2) scale is designed according to selection of response choices and writing instructions;

3) initial version of scale is pilot-tested with a small number of respondents who are asked to critique the scale. They usually indicate which items are ambiguous or confusing, and which items cannot be evaluated along the dimension chosen;

4) a full administration and item analysis is conducted. A sample of 100 to 200 respondents complete the scale;

5) the scale is validated and normed. The norm describes the distributional characteristics of a given population on the scale. Individual scores on the scale are then interpreted in relation to the distribution of the scores in the population.

3. RELIABILITY IN A VIEW OF CLASSICAL TRUE-SCORE THEORY OF MEASUREMENT

Classical True-Score Theory (CTM) describes to what extent errors of measurement can influence observed scores (1904 a-b, 1907; Guliford, 1936; Gulliksen, 1950). Typically in CTM true scores represent an average score taken over repeated independent measures with the same score as a

theoretical idea. The formula of CTM is expressed as follows (Spearman 1904 a-b, 1907; Guliford,1936; Gulliksen, 1950):

$$X = T + E, \tag{4.0}$$

where: X – observed score, T – true score, E – error score or error of measurement.

R**eliability** as a word has a positive connotation. For anything to be characterized by a person as reliable is to be described in positive terms. So it is with any type of test (scale development), experiment, or measuring procedure application. If it is reliable (e.g. scale), then it has gone a long way toward gaining scientific acceptance. Nunnally (1978) defined reliability as *"the extent to which measurements / tests are repeatable and that any random influence which tends to make measurements different from occasion to occasion, is the source of error".*

In order to formally define reliability, we ought to establish some notation for the basic concepts derived from Classical True-Score Theory of Measurement (CTM). Conceptually, the true score is a perfect measure of the property being measured. However, in practice, the true score can never really be known and generally is assumed to be the mean score of a large number of administrations of the same scale to the same subject. The fundamental equation (4.0) of the CTM is, when true scores and error scores are assumed to add (rather than to have some other relationship, such as a multiplicative one). If then two measurements have observed scores X and X' that satisfy the assumptions ranging from 1 to 5 (Allen and Yen, 1979):

$X = T + E,$

1. $E(e) = 0,$ the expected value (population mean) of error scores for any person is 0,
2. $\rho_{ET} = 0,$
3. $\rho_{E_1E_2} = 0,$
4. $\rho_{E_1T_2} = 0.$

and if for every population of persons, $T = T'$ and $\sigma_E = \sigma'_E$, then these measurements are called **parallel measurements**. And if two measurements $X_1 = T + E_1$ and $X_2 = T + E_2$ are said to be parallel, we assume that $\sigma^1_{E_1} = \sigma^2_{E_2}$ and $\rho_{E_1E_2} = 0.$ It then follows that $\sigma^1_{X_1} = \sigma^2_{X_2} = \sigma^2_X$. Omitting the

subscripts from X_1 and X_2 we obtain their correlation coefficient denoted as (Allen and Yen, 1979):

$$\rho_{tt} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2. \tag{4.1}$$

where: ρ_{tt} denotes the population correlation coefficient between parallel measurements (population reliability coefficient).

The correlation between the parallel measurements gives the possibility to estimate the reliability, but it requires quite rigorous assumptions on the measurement errors that take place during process of measurement. Usually the correlation coefficient yields 1.0 if two distributions for both test are equal or parallel. Theoretically parallel measurements (tests) ought to have the same average variance and correlation between those pairs of measurements.

If two measurements on the other hand have observed scores X_1 and X_2 that satisfy the above assumptions 1 through 5, and if for every population of persons, $T_1 = T_2 + c$, where c denotes a constant, then these measurements are called **essentially *tau*-equivalent measurements.** This allows for separation of the true score variance from the measurement error variance and the variance of observed score X which can be written as follows (Allen and Yen, 1979):

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{4.2}$$

Having returned to parallel measurement principles, we can define reliability according to CTM theoretical model (4.0) that is the **squared correlation** between X and T (true score) denoted as (Allen and Yen, 1979):

$$\rho_{tt} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} = \frac{1}{1 + \frac{\sigma_E^2}{\sigma_T^2}}. \tag{4.3}$$

The definition gives **three equivalent forms of reliability,** expressed with the components of Eq. (4.3). The first form will tell us that reliability is the ratio of the true score variance σ_T^2 to the total variance σ_X^2. The second one expresses the measurement error variance σ_E^2. The last form of the definition, which does not contain σ_X^2 explicitly follows by dividing one by

the inverse of the first form. Observed scores at best, account for, predict, or explain true scores when this linear correlation ratio (Allen and Yen, 1979):

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}. \tag{4.4}$$

is 1.0 – that is when all observed score variance is true score variance. 1.0 minus the ratio of error score variance to observed score variance gives us an alternative way of understanding or interpreting ρ_{XT}^2.

Differences that arise between different types of measurement such as: parallel, tau-equivalent and congeneric are due to some few facts (Graham, 2006). Namely in **parallel measurement** the amount of variation in the item score that is determined by the true score is the same for all items. In fact this implies that the expected value of each of the items will be the same. The easiest practical example one can imagine of something like this would be a situation where one employs the exactly same measure on something on multiple occasions where one has no reason to expect any kind of testing effect or change in the true score over the period in which the multiple measures were administered. The parallel measurement is the most restrictive measurement for use in defining the composite true score.

Tau-equivalent measurement is identical to the more restrictive parallel model, save that individual item error variances are freed to differ from one another. This implies that individual items measure the same latent variable on the same scale with the same degree of precision, but with possibly different amounts of error. The **essentially tau-equivalent measurement** is, as its name implies, essentially the same as the tau-equivalent model. Essential tau-equivalence assumes that each item measures the same latent variable, on the same scale, but with possibly different degrees of precision. Again, as with the tau-equivalent model, the essentially tau-equivalent model allows for possibly different error variances. The difference between item precision and scale is an important distinction to make. Whereas tau-equivalence assumes that the items true scores are equal across items, the essentially tau-equivalent measurement allows each item's true score to differ by an additive constant unique to each pair of variables.

Finally the **congeneric measurement** is the least restrictive, most general measurement of use for reliability estimation. The congeneric assumes that each individual item measures the same latent variable, with possibly different scales, with possibly different degrees of precision, and with possibly different amounts of error. Whereas the essentially tau-equivalent model allows item true scores to differ by only an additive constant, the

congeneric model assumes a linear relationship between item true scores, allowing for both an additive and a multiplicative constant between each pair of item true scores.

4. FACTORS AFFECTING RELIABILITY LEVEL AND METHODS OF RELIABILITY ESTIMATION IN CTM

According to Symonds (1928) there are several factors in Classical True-True Score Theory (CTM) which affect considerably the reliability level in the measurement. Six factors are also general considerations in scale construction and they are related to:

- **the number of items** – reliability increases as the number of items in a scale increases.
- **the range of item difficulty** – the narrower the range of item difficulty, the greater the reliability of the scale. Items that are answered correctly (or incorrectly) by all individuals do not contribute to variability within a test (measurement) and decrease the number of functional items.
- **evenness in scaling** – the result of a developing scale with items at the same level of difficulty is equivalent to reducing the number of items. All items of equal difficulty should be answered either correctly or incorrectly. The extreme case is with two sets of items: 1) those answered correctly by all persons, and 2) those answered incorrectly. This situation reduces the test to two items. Optimally, the test will be evenly scaled across a range of item difficulties.
- **interdependence of measured items** – lower estimates of the reliability will be achieved if the answer to one item is suggested by another item, or if the meaning of one item is dependent upon a previous item.
- **guessing** – scale reliability decreases as the likelihood of guessing the correct answer increases.
- **homogeneity** – if items of the measurement have different concepts, then the scale reliability will decrease.

Reliability estimation for unidimensional reflective indicators

There are different methods of reliability estimation when unidimensional reflective indicators are considered in particular. However, we focus here only on the most common and widely practiced solutions. They are presented in Table 1.

In **determining reliability estimates**, test-retest or parallel forms estimates should be used, because most of the internal consistency measures would be inaccurate. The use of coefficient Alpha or Kuder-Richardson would produce a *lower bound* for a test's reliability. The lower bound equals the test reliability if the components in the test are essentially tau-equivalent. Coefficient Alpha and the Kuder-Richardson formulas should be used only for homogeneous tests, since they basically reflect item homogeneity.

If the test measures a variety of traits, coefficient Alpha and the Kuder-Richardson reliability will be inappropriately low. That is why the Kuder-Richardson formula 20 (KR20) gives good level of reliability of a test where components are dichotomous items. And Kuder-Richardson 21 (KR1) equals the test reliability if the dichotomous items in the test have equal item difficulties.

The Spearman-Brown formula can overstimate or underestimate a test's reliability if the components of the test are not parallel. When the components of a test are parallel, the Spearman-Brown formula is very useful for judging the effects that changes in test length on reliability. In short it is useful for estimating the reliability of a test with altered length and it offers reasonable estimates if the test length is changed by adding or omitting parallel versions of the original test items. Since reliability tends to be lower for shorter tests (Allen and Yen, 1979).

As far as the split-halves approach is concerned, the major problem is that the correlation between halves will differ somewhat depending on how the total number of items is divided into these halves. In contrast coefficient Alpha (that is typically used in multi-item scale, e.g. summated scale) is particularly easy to use because it requires only a single test / measurement administration. Moreover, it is a very general reliability coefficient, encompassing both the Spearman-Brown prophecy formula as well as the Kuder-Richardson 20. Alpha is also easy to compute, especially if one is working with a correlation matrix. Minimal effort that is required to compute Alpha is more than repaid by the substantial information that it conveys about the reliability of a scale (Carmines and Zeller, 1979).

Table 1

Selected measurement reliability methods

Internal consistency	Function
Kuder -Richardson 20/21	Used for true score theory approach (raw scores) Used for dichotomous responses
Cronbach's Alpha	Used for true score theory approach (raw scores) Used for polytomous responses
Spearman-Brown formula	Allows calculation for hypothetical reliabilities
Test-Retest Test-Retest correlation	Used when same respondents are measured again
Parallel-Test Alternate form correlation	Used when two measurements are parallel, e.g. take time at the same moment

Source: own construction based on Wilson, 2005

Test-Retest method

The reliability obtained by repeated administration is referred to as test-retest reliability. The test-retest method remains the most popular one to estimate the *stability of a test (measurement) over time.* But this method is vulnerable to artifact from random variability in responses, changes in the individuals taking the measurement, and differences in the method by which measurements (tests) are administered.

The aim of this method is to investigate the variation in the items locations due to the instrument, not due to real change in respondents' locations. Both measurements (first test and second) should be close enough together to assume that there has been little real change. The equation for the two tests may be found in Carmines and Zeller (1979) and also Magnusson (1981).

This method will work better where a stable latent variable (constructed scale) is measured with forgettable items, as compared with a less stable latent variable measured with memorable items. Moreover between first survey (test) and second survey (retest) there is no strictly defined interval time, but usually, a second survey is conducted straight away after the first one is finished. If there would be a short time interval, it might in consequence cause **a carry-over effects** due to respondents' memory, practice, or mood more likely. On the other hand, a long interval would make effects due to changes in information or moods likely too. In consequence when choosing test-retest method, researchers are often only able to obtain a measure of a phenomenon at a single point in time. Not only can it be unduly expensive to obtain measurements at multiple points in time but it can be impractical.

In the test-retest approach, the computed correlations and their interpretation are not straightforward. Naive interpretation of test-retest correlations can drastically underestimate the degree of reliability in measurements over time by interpreting true change as measurement instability. A low test-retest correlation may not indicate that the reliability of the test is low, but may instead signify that the underlying theoretical construct itself has changed.

Parallel-Test method

In this type of method if researcher cannot provide the same or at least comparable score level within *test* and *retest,* then one can choose *parallel-test.* The two alternate copies of the instrument are then administered and calibrated, and then two sets of locations are correlated to produce alternate forms reliability coefficient. This method can be performed provided the following criteria of measurement will be met (Brzezinski, 2007):

- equal scores on averages between first and second test,
- equal variances,
- equal intercorrelations for each measured item on two tests.
- no frame of time between two tests. Second test comes right away, after the first one is performed and finished.

Split-halves method

In this type of method, reliability estimation is calculated first when one prepares a set of items for a given scale. In the next stage, the same set of items is divided into two subsets. Thirdly, subsets are correlated with each other in order to yield suitable reliability level. The whole test is divided into two parts, with alternate forms of each other. If the halves of the test are parallel, the reliability of the whole test is estimated using the Spearman-Brown formula. But if the halves are essentially *tau*-equivalent, then coefficient α is computed. Allen and Yen (1979).

There are several **ways of splitting a set of items.** One is called *first-half last-half,* where items interact with each other and thus affect each subset. This is the case when items are scattered throughout a lengthy questionnaire and where respondents might be more fatigued when completing the second half of the scale (DeVellis, 2003). In another *odd-even reliability,* we have a subset of odd-numbered items that is compared to the even-numbered items. In order to split the halves properly, one needs to sort and rank items according to their level of difficulty. This process of items extraction is as follows:

A) = 1, 3, 5, 7 – as part of odd-numbered items,

B) = 2, 4, 6, 8 – part of even-numbered items.

Later on we calculate the correlation coefficient between total scores of two "halves" tests, and estimate reliability of test, applying for it Spearman-Brown formula. While splitting particular tested items into two halves, an equal level of variance on both halves must be assumed (DeVellis, 2003).

The other alternative types of reliability measurement for split-half scale development are: balanced halves and random halves. In the former, one must identify some potentially important item characteristics (such as item length, or type of response indicating presence or absence of the variable in question). The two halves are then formed, so the characteristics of both of them are equally represented either in the first or second half, each according to the same level of items word-formations and so on. In contrast, through random halves one obtains halves based on random allocation of each item within one of the two subsets that should be eventually correlated with each other. The quality of this approach depends on the number of items chosen for analysis, the number of characteristics of subject analysis and degree of independence among items as well.

Kelly (1958) advocated improving split-half reliability by making the split tests as similar as possible through matching of item content and difficulty. Cronbach (1951) believed that random splits would yield

coefficients lower to parallel form (or planned split). Finally the split-half method is probably one of the best methods for estimating reliability where test (measurement) answers will be corrected for guessing, or where item weighting will be used.

Internal consistency reliability methods

Internal consistency reliability methods are based on statistical exploration of items either in covariance or correlation with total test score *in one time approach*. *Internal consistency* refers to the interrelatedness of a set of items. Sometimes it is confused with homogeneity as though they were synonymous. *Homogeneity* refers to the unidimensionality of the set of items. Although internal consistency is certainly necessary for homogeneity, it is not sufficient (Hattie, 1985; Cortina, 1993).

Cronbach (1951) viewed reliability (including internal consistency measures) as the proportion of test variance that was attributable to group and general factors. Specific item variance, or uniqueness, was considered an error. Also in Cronbach's point of view, Alpha will be an underestimate of reliability (as he defined it) unless the inter-item correlation matrix will be of unit rank (i.e. unidimensional). Moreover Cronbach's early statements (1947) about reliability suggest that the reliability of a multidimensional measure can only be estimated by correlating scores on parallel forms of a test that each represent the same structure.

- Cronbach Alpha

Alpha coefficient is the general version of the Kuder-Richardson 20 coefficient of equivalence. It is a general version, because the K-R coefficient applies only to dichotomous items, whereas Alpha applies to any set of items regardless of the response scale. Cronbach's coefficient Alpha is based on the assumption of compound symmetry (equal item variances and covariances) as far as items reliability estimation is concerned. Specifically Alpha coefficient is (Cronbach, 1951):

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum_{i=1}^{k} \sigma_{X_i}^2}{\sigma_X^2} \right). \tag{5.0}$$

where: k – number of items in the scale, where $k \geq 2$, $\sigma_{X_i}^2$ – variance of i-th item, σ_X^2 – total variance of the scale.

If there is no true score but only error in the items (which is esoteric and unique), then the variance of the sum will be the same as the sum of variances for individual items. In consequence, coefficient Alpha will equal to zero. If all items are perfectly reliable and measure the same thing (true score), then coefficient Alpha will equal to 1 (Peter, 1979; Netemeyer et al., 2003).

And as far as the **Alpha's standard error** is concerned it is inversely related to sample size. Thus, researchers seeking to improve the predictive ability of their scales can do so indirectly through increasing sample size. In case of **scale length**, it has both direct (via its influence on Alpha) and indirect (via its influence on the standard error) measurement benefits to the researcher. Finally, the effects of **item inter-correlations** on both the standard error and Alpha are dramatic, e.g. stronger correlations among the items drastically reduce the standard error and increase Alpha.

Research findings also confirm that larger heterogeneity within the covariance matrix negatively impacts reliability. Specifically, it decreases the precision of the Alpha estimate. Some analysis provides also insights for considering **additional new items to a scale.** It is not always that Alpha will be enhanced when more items will be added. It depends rather on the length of the original scale and their items correlations. The items added must be of increasingly high quality (in terms of their correlations with the original items) to improve Alpha at all (Iacobucci et al., 2005).

- Kuder and Richardson (K-R 20 and K-R 21),

Kuder and Richardson (1937) introduced another approach to reliability measurement, which required only one test administration. They proposed a method for estimating test reliability and its hypothetical equivalent. They believed their estimate would be applicable to any unidimensional test / scale measurement where items were unit weighted. Inter item coefficients were allowed to vary between their possible limits, as were varying proportions of correct answers. Items did not need to be equally difficult or equally correlated with other items. In **K-R 20** approach instead of using two tests forms, the actual test was compared with its hypothetical equivalent. Because reliability had been defined as the solution between two forms of a test, the coefficient reliability could be estimated by computing the correlation between the actual test and its hypothetical counterpart.

For items that are scored dichotomously, a proposed formula enabled to split the test, consisting of k items into k parts. That means, one part of it equals to the other one, particular item. The analysis should be based on parallel test items. For example, for answers 0 or 1 with equal level of

difficulty, a fraction of good answers p equals the fraction of bad answers q. If one doesn't know the level of difficulty on particular tested items (as happens in *K-R 20* formula), then we can accept *K-R 21* where this level is estimated approximately, or is comparable among test items. Both formulas are widely discussed in Kuder and Richardson (1937) and also in Ferguson and Takane (2009).

Kuder and Richardson in their investigations emphasized that K-R 20 required item variances, whereas K-R 21 required item difficulties. If items were equally difficult, then two values would be the same. Otherwise, the reliability estimate from K-R 20 would exceed that from K-R 21.

Reliability estimation for multidimensional reflective indicators

If there are more different types of scales under investigation, then their sets of items must be assigned separately to each measured subscale. Next (for each of subscale), a reliability estimation is calculated. The problem of many items and sub scales reliability estimation was solved by Armor (1974), when he proposed Theta and Omega reliability formulas.

Theta coefficient can be easily understood once we consider properties of Principal Components, i.e. the Factor Analysis Model on which this reliability coefficient is based. As a result it depends on whether we measure a single phenomenon or more than just one. In a case when a set of items is measuring a single phenomenon we could assume that in principal components analysis: 1) the first extracted component should explain a large proportion of the variance in the items (e.g., > 40%), 2) subsequent components should explain fairly equal proportions of the remaining variance except for a gradual decrease, 3) all of most the items should have substantial loadings on the first component (e.g., > 0.3), and 4) all or most of the items should have higher loadings on the first component than on subsequent components.

In contrast, if we measure many phenomena, principal components analysis of items should meet the following conditions: 1) the number of statistically meaningful components should equal the number of hypothesized phenomena, 2) after rotation, specific items should have higher factor loadings on the hypothesized relevant component than on the other components, and 3) components extracted subsequently to the number of hypothesized components should be statistically unimportant and substantively uninterpretable (Carmines and Zeller, 1979).

Once the items and their corresponding weights are chosen, the reliability of the resulting scale can be estimated using the following formula for Theta (Armor, 1974; Carmines and Zeller, 1979):

$$\theta = \frac{k}{k-1}\left(1 - \frac{1}{\lambda_1}\right). \tag{5.1}$$

where λ is the highest eigenvalue of correlation matrix among all the items of the scale.

Other coefficient useful for reliability estimation is called Omega Ω and is applied in Factor Analysis (Armor, 1974; Carmines and Zeller, 1979). It is expressed as follows:

$$\Omega = 1 - \frac{\sum_{i=1}^{k} \sigma_{X_i}^2 - \sum_{i=1}^{k} \sigma_{X_i}^2 h_i^2}{\sum_{i=1}^{k}\sum_{j=1}^{k} \mathrm{Cov}\left(X_i, X_j\right)}. \tag{5.2}$$

where: h_i^2 – denotes communality of observable variable X_i estimated on correlation matrix.

This reliability estimate is applied for linear latent variables – subscales. Omega is based on the Common Factor Analysis Model (CFAM), where unities are replaced by communality estimates in the main diagonal of the correlation matrix prior to factoring.

Between Theta θ and Omega Ω there are two important differences. First, they are based on different factor-analytic models. Theta reliability coefficient Theta θ is grounded in the Principal Components Model, whereas Omega Ω is based on CFAM. This means that one always uses 1.0's in the main diagonal to compute the eigenvalues on which Theta is based but the value of Omega depends, in part, on communalities, which are estimated quantities not fixed ones. Because Omega is based on estimated communalities, there is an element of indeterminacy in its calculation that is not present in Theta.

Second, unlike Theta, Omega does not assess the reliability of separate scales in the event of multiple dimensions (Armor, 1974). Omega rather provides a coefficient that estimates the reliability of *all the common factors* in a given item set (Carmines and Zeller, 1979).

5. CONFIRMATIVE UNIDIMENSIONAL AND MULTIDIMENSIONAL FORMATIVE INDICATORS

Confirmative Factor Analysis (CFA) in a view of Classical Theory of Measurement allows on the other hand to perform an extended analysis of the reliability and estimation of additional coefficients. Usually it is assumed that each variable on observable one $X_i\left(i=1,...,k\right)$, is affected only by one latent variable (Kozyra, 2004):

$$X_i = \lambda_i \xi + \delta_i. \tag{6.0}$$

where: λ_i – coefficient between observable variable depending on latent variable,

ξ – latent variable,

δ_i – random errors in the measured latent variable,

Reliability coefficient of single indicator X_i is therefore given by (Bollen, 1989):

$$\rho_{X_i X_i} = \frac{\lambda_i^2 \sigma_\xi^2}{\sigma_{X_i}^2} = \frac{\lambda_i^2 \sigma_\xi^2}{\lambda_i^2 \sigma_\xi^2 + \theta_{ii}}. \tag{6.1}$$

where – θ_{ii} – measurement errors.

This equation may be simplified in case of standardized latent variable ξ:

$$\rho_{X_i X_i} = \frac{\lambda_i^2}{\lambda_i^2 + \theta_{ii}}. \tag{6.2}$$

The reliability coefficient of the unweighted sum of *k* indicators $X = X_1 + ... + X_k$ will be then described by (Bollen, 1989):

$$\rho_{XX} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2 \sigma_\xi^2}{\sigma_X^2}. \tag{6.3}$$

In the denominator of this equation there is a sum of all elements of the covariance matrix of variables, $X_i\left(i=1,...,k\right)$, whereas in the numerator, this sum is deducted, e.g. the variances of measurement errors occurring in the covariance matrix. For the standardized latent variable ξ and

uncorrelated with each measurement errors θ_{ii}, this equation is simplified. As a result we obtain (Bollen, 1989):

$$\rho_{XX} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\sigma_X^2} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right) + \sum_{i=1}^{k} \theta_{ii}}. \tag{6.4}$$

And this equation is one of the most commonly applied in the **reliability estimation for unidimensional indicators,** which are considered in Confirmation Factor Analysis (CFA). If its value is obtained on the 0.7 level (Hair et al., 1992), then reliability measurement is strongly satisfying.

Another quite often practiced reliability coefficient is the variance extracted coefficient, which for standardized latent variables is expressed as follows (Bollen, 1989):

$$\sigma_{E_X}^2 = \frac{\sum_{i=1}^{k} \sigma_{X_i}^2 - \sum_{i=1}^{k} \theta_{ii}}{\sum_{i=1}^{k} \sigma_X^2} = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{k} \lambda_i^2 + \sum_{i=1}^{k} \theta_{ii}}. \tag{6.5}$$

If this coefficient reaches the level of at least 0.5 (Hair et al, 1992), then the estimation of reliability measurement is sufficient. In the case when measurement errors of observable variables are correlated, then the formula of reliability takes the following form (Bollen, 1989):

$$\rho_{XX} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} \theta_{ii}}. \tag{6.6}$$

On the other hand if latent variable is measured on weighted sum of observable variables such as: $X_i = wX_1 + ... + w_k X_k$, then reliability coefficient for standardized latent variable and correlated measurement errors will be expressed as follows (Bollen, 1989; Kozyra, 2004):

$$\rho_{XX} = \frac{\left(\sum_{i=1}^{k} w\lambda_i\right)^2}{\left(\sum_{i=1}^{k} w\lambda_i\right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} w_i w_j \theta_{ii}}. \tag{6.7}$$

And if measurement errors remain uncorrelated, then the above equation is simplified to (Bollen, 1989; Kozyra, 2004):

$$\rho_{XX} = \frac{\left(\sum_{i=1}^{k} w\lambda_i\right)^2}{\left(\sum_{i=1}^{k} w\lambda_i\right)^2 + \sum_{i=1}^{k} w_i^2\theta_{ii}}. \tag{6.8}$$

At given variance values for uncorrelated measurement errors, the highest level of reliability for the weighted sum of observable variables, will be obtained provided the weights are equal (Ostasiewicz, 2002, 2003), e.g., $w_i = \frac{\lambda_i}{\theta_{ii}}$.

As a result the following reliability coefficient is obtained (Bollen, 1989; Kozyra, 2004; Ostasiewicz, 2002, 2003):

$$\rho_{XX} = \frac{\left(\sum_{i=1}^{k} \frac{\lambda_i^2}{\theta_{ii}}\right)^2}{\left(\sum_{i=1}^{k} \frac{\lambda_i^2}{\theta_{ii}}\right)^2 + \sum_{i=1}^{k} \frac{\lambda_i^2}{\theta_{ii}}} = \frac{\sum_{i=1}^{k} \frac{\lambda_i^2}{\theta_{ii}}}{\left(\sum_{i=1}^{k} \frac{\lambda_i^2}{\theta_{ii}} + 1\right)}. \tag{6.9}$$

The other well-known index such as Dillon-Goldstein's (or Jöreskog's) rho (Wertz et al., 1974) better known as *composite reliability* is considered as homogenous if it is larger than 0,

$$\rho = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \left(1 - \lambda_i^2\right)}. \tag{6.10}$$

Dillon-Goldstein's rho is considered to be a better indicator than Cronbach's alpha. Indeed, the latter assumes the so-called tau-equivalence (or paralleled) of the manifest variables, i.e. each manifest variable is assumed to be equally important in defining the latent variable. Dillon-Goldstein's rho does not make this assumption as it is based on the results from the model (i.e. the loadings) rather than the correlations observed between the manifest variables in the dataset.

All the above described reliability coefficients are applied only when observable variable compose single latent variable. **In the case when observable variables are affected with more than just one latent variable,** then reliability coefficient is estimated with squared multiple correlation $R^2_{X_i}$ where X_i is crossed in the analysis with latent variables $\xi\ (j = 1, \ldots, n)$. The coefficient is given by the formula (Bollen, 1989):

$$R^2_{X_i} = 1 - \frac{\theta_{ii}}{\sigma^2_{X_i}}, \tag{6.11}$$

where: $\sigma^2_{X_i}$ denotes variance of X_i due to estimated model. This type of reliability coefficient for joint measurement model is likely to be described by analogy as coefficient for observable variables (Bollen, 1989):

$$R^2_{X_i} = 1 - \frac{\det\left(\Theta_\delta\right)}{\det\left(\Sigma_{XX}\right)}. \tag{6.12}$$

This coefficient enables the researcher to indicate what part of the total observable variables variance is explained by latent variables in a confirmatory model.

6. ITEM RESPONSE THEORY (IRT)

Item Response Theory (IRT) is an alternative option to Classical True-Score Theory of Measurement (CTM). It originates from the early works of Lazarsfeld, Green and Torgerson conducted in the 1950s (Aranowska, 2005). However, since Lord and Novick's (1968) classic book introduced model-based measurement, a quiet revolution has occurred in measurement theory. Item Response Theory (IRT) is known also as Latent Trait Theory, in which trait level estimates depend on both persons' responses and on the properties of the items that are administered. This theory breaks off from the fundamental constriction of CTM, namely the assumption of parallel items and dependence of the reliability in measurement within the characteristics in the sample. In contrast to the parallel items position in a scale associated with model of CTM, in IRT theory one needs to describe systematic relationships between items, which form a hierarchical monotone relationship. An example of such a scale is a well-known monotonic Guttman scale, whose hierarchical and monotonic nature reflects items on scale measuring attitudes e.g. towards product or some type of service. This simplified scale consists no longer of parallel item positions as compared to the Likert scale (in CTM), but its items are strongly hierarchical. The unidimensional **Guttman** scale is usually not a unidimensional scale as it is in CTM in context of factorial approach. In the case of the Guttman scale, the underlying factor of the latent scale is hidden as a cause of differences in the responses to scale items. As a result, there occurs a strong correlation between adjacent items of Guttman scale, and there are differences in the average values of results for specific items. In the case of

the unidimensional Guttman scale application of factor analysis will not lead to the disclosure of one common factor for this scale (Sagan, 2002).

In literature there are many diverse IRT models. We do not discuss them all here. As a matter of fact the main focus was put on **unidimensional models.** The origins of IRT and early models emphasized dichotomous item formats (e.g. the Rasch model, 1960), which were further extended to other item formats, such as rating scales (Andrich, 1978, 1982) and partial credit scoring (Masters, 1982). Next, the undimensional IRT models have been generalized to **multidimensional models** where traits could be measured and scales could be constructed by comparisons within tasks (Kelderman and Rijkes, 1994), changes across conditions (Embretson, 1991), subtasks representing underlying cognitive components (Embretson, 1984), or conditioning on measurement-taking strategy (Rost, 1990).

As we can infer, IRT represents a family of models rather than a theory specifying a single set of procedures. One important way in which the alternative IRT models differ is the number of item parameters with which they are concerned. It usually concentrates on three aspects of an item's performance. They are: **items difficulty, items capacity to discriminate, items susceptibility to false positives.**

Item Response Theory in its underlying assumptions is grounded in probability area, where respondent has a potential ability Θ and probability equal of 1, to give the correct answer on i-th item (under investigation in the measurement), and when the difficulty level is lower or equal to Θ. Otherwise the success of making right answer is 0. According to Torgerson (1958) there is function f of such probability, that is expressed as follows:

$$P_i(\Theta) = a_i\Theta + C_i. \tag{7.0}$$

where: a_i denotes coefficient of slope line.

In fact, IRT assumes a probabilistic model wherein the likelihood that individuals will respond in a particular manner to an item or question is proportional to that individual's position on a latent trait or continuum. This can be also illustrated on Figure 1, according to ICC (Item Characteristic Curve). This curve or relation in the model was deterministic as far as Guttman solution was concerned. The latest models have a probabilistic nature. In Rasch's model, or Birnbaum's model, it is a parametric logistic function. ICC typically describes how changes in trait level relate to changes in probability of a specified response. For dichotomous items, in which a specified response is considered "correct" or in "agreement" with an item,

the ICC regresses the probability of item success on trait level. For polytomous items, such as rating scale, the ICC regresses the probability of responses in each category on trait level.

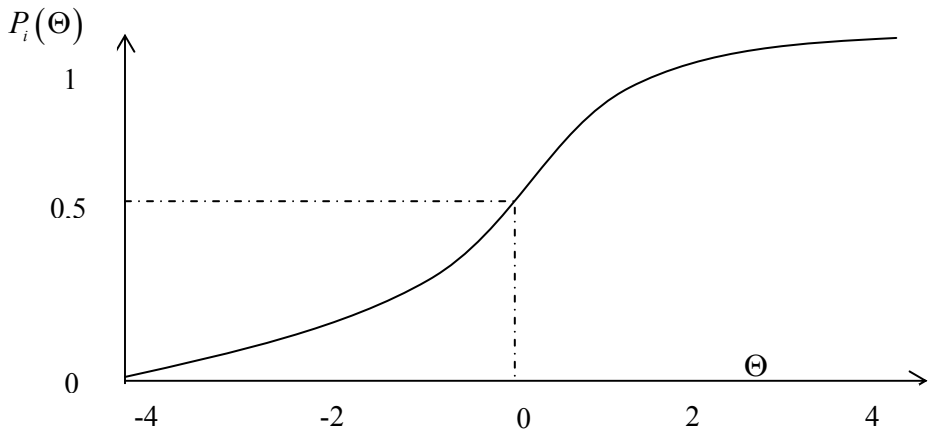

Figure 1. Item characteristic curve

Source: based on: Rosenbaum 1987, Raju 1988, Embretson and Reise 2000; Aranowska, 2005

Because probability cannot be greater than 1 or lower than 0, there were introduced other modeling solutions such as **Logit Probability** (Aranowska, 2005):

$$P_i(\Theta) = \log \frac{P_i(\Theta)}{1 - P_i(\Theta)}. \tag{7.1}$$

Logit fraction expresses ratio of "success" to "failure". Logarithm function transforms positive numbers into real axis. That is, when "success" is greater than "failure", positive numbers appear. On the other hand, when "success" is lower than "failure" – into negative. When the fraction equals unity – into 0.

An early and still popular member of the IRT family is **Rasch Model,** which quantifies the difficulty parameter. Rasch understood that the relationship between above transformation and Θ should be expressed in the following way (Rasch, 1960):

$$\log \frac{P_i(\Theta)}{1 - P_i(\Theta)} = \Theta - b_i. \tag{7.2}$$

where: b_i – parameter corresponding to the position of *i*-th item in the measurement, assuming that difficulty item is following Θ;

$\Theta - b_i$, when "success" of correct answer for *i*-th position is equal 0.5 where answer is random.

For example, if the odds that a person passes an item is 4/1, then out of five chances, four successes and one failure are expected. Alternatively, odds are the probability of "success" divided by the probability of "failure", which would be 0.80/0.20 in this example. If the trait level equals item difficulty, then the *log odds* of "success" will be zero. Taking the antilog of zero yields and odds of 1.0 (or 0.50/0.50), which means that a person is as likely to succeed as to fail on this particular item (Embretson and Reise, 2000). Solving next the equation we obtain:

$$P(\Theta)_i \frac{e^{(\Theta - b_i)}}{1 + e^{(\Theta - b_i)}} = \frac{1}{1 + e^{-(\Theta - b_i)}}. \tag{7.3}$$

This model reflects the principle for **scaling of unidimensional metrical trait** (variable). Therefore it may be also applied for interval scale construction. Moreover in this equation (known as one-parameter logistic) dependent variable is predicted as a probability rather than as a log odds. This is due to its exponential form in predicting probabilities and to the inclusion of only one item parameter (e.g. difficulty) to represent item differences.

In the course of time, the IRT basic Rasch Model was extended with additional parameters. As a result there appeared:

- **two-parametric logistic model** of Birnbaum's (1968):

$$P(\Theta)_i \frac{e^{Da_i(\Theta - b_i)}}{1 + e^{Da_i(\Theta - b_i)}} = \frac{1}{1 + e^{-Da_i(\Theta - b_i)}}. \tag{7.4}$$

where: a_i denotes parameter of additional item discrimination, corresponding to curve slope;

D – some constant.

when $D = 1,7$ then (7.4) reflects normal distribution function.

In the two parameter logistic model, item discrimination is included. The model has two parameters to represent item properties. Both item difficulty and item discrimination are included in the exponential form of logistic model. Item discrimination is a multiplier of the difference between trait level and item difficulty. Item discriminations are related to the biserial correlations between item responses and total scores. Hence for this equation

(based on two-parameter logistic solution), the impact of the difference between trait level and item difficulty depends on the discriminating power of the item. Specifically, the difference between trait level and item difficulty has a greater impact on the probabilities of highly discriminating items (Embretson and Reise, 2000).

- **three-parameter logistic model** supplemented with guessing parameter (Birnbaum, 1968):

$$P(\Theta)_i = c_i + (1-c_i)\frac{e^{Da_i(\Theta-b_i)}}{1+e^{Da_i(\Theta-b_i)}} = c_i + \frac{1-c_i}{1+e^{-Da_i(\Theta-b_i)}}, \text{ for } i=1,...,k, \quad (7.5)$$

where: k – denotes number of item position.

In three-parameter solution, a one more parameter is added to represent an item characteristics curve that does not fall to zero. For example, when an item will be solved by guessing, as in multiple choice cognitive items, the probability of "success" is substantially greater than zero, even for low trait levels. This model accommodates guessing by adding a lower asymptote parameter c_i.

Last, unlike the model of Rasch and Birnbaum, we have also in IRT family models, a non-parametric Mokken model (stochastic model), where in the constructed scale, the respondent is giving positive answers on items that have probability of positive response to the less difficult items significantly higher than zero (Mokken, 1971).

For assessing the degree of monotonicity of scale within the meaning of Guttman scale, we are allowed to use many methods. The most important is the coefficient of scalar and reproducibility scale. There are also used chi-square test of significance on Guttman scale. The scalar coefficient is given by (Sagan, 2002):

$$C_S = 1 - \frac{E}{X}. \quad (7.6)$$

where: E – number of errors in Guttman table (constructed based on item responses of the respective scale),

X – number of accidental errors of the scale, expressed by:

$$X = p(n - T_n). \quad (7.7)$$

where: p – probability of giving response on the respective item (0.5 for response such as "yes" – „no"),

n – number of choices in items,

T_n – number of elections in the largest category for each item.

The Guttman scale is a monotonic scale (reliable) if scalar ratio is greater than 0.6. Coefficient of reproducibility is an alternative indicator of monotonicity scale:

$$C_r = 1 - \frac{E}{N}. \tag{7.8}$$

where: E – number of errors in Guttman's table,

N – number of all the choices on the scale. It is a product of the number of items and number of respondents.

Guttman scale is a monotonic scale (reliable), if the coefficient of reproducibility is greater than 0.9. Also synthetic indicator of monotonicity (reliability) in the scale is H-Loevinger' coefficient:

$$H = 1 - \frac{E}{E_o}. \tag{7.9}$$

where: E – probability of errors in in a given Guttman table,

E_o – probability of errors for a completely independent position in the scale.

For a reliable scale, a minimum value of H-Loevinger should be greater than 0.3, and the scale should have strong values of H greater than 0.5. In the case of monotonic items or various scale difficulty, H coefficient is a better measure of the reliability of the scale than the α-Cronbach.

7. DIFFERENCES BETWEEN CLASSICAL TRUE-SCORE THEORY (CTM) AND ITEM RESPONSE THEORY (IRT)

In CTM, true score estimates are typically obtained by **summing responses across items** which in the next phase form a respective scale. In IRT, estimating trait levels involves a search process for optimal estimates. Also in IRT one can construct a scale but its items entities have contrary (to CTM) structure of responses distribution, e.g., they are based on dichotomous answers (Mokken, 1971). In both approaches, measurement is based on latent variables (or extracted scales). The particular item response or test score is defined as **indicator** of a person's standing on the latent variable, but it does not completely define the latent variable. Typically all marketing research measurement, e.g. customers (based on human internal traits) are usually of indirect characteristics. In CTM the independent variables are combined additively and directly to predict the dependent

variable. In IRT (known as a strong modeling method) strong assumptions must be met in advance.

The comparison of the Classical True-Score Theory of Measurement (known as "old rules") and Item Response Theory ("new rules") could be started from Lord and Novick (1968) early derivations. New rules of measurement IRT, are fundamentally different from the old rules – CTM. Many old rules, in fact, must be revised, generalized, or even abandoned. Hence a description is here introduced concerning the differences between two measurement concepts. Some of them are summed up in Table 2.

In first aspect – **standard error of measurement,** difference appears on whether the standard error of measurement is constant or variable among the scores in the same population. In CTM at standard error of measurement, a constancy is specified, whereas in IRT a variability is considered. Besides, measurement is different whether the standard error is specific or general across populations. In CTM it is rather population specific, whereas IRT it is population general. Moreover if we estimate standard error in IRT we assume that the relationship between trait score and raw score is nonlinear, and the confidence interval band becomes increasingly wide for extreme scores. Unlike CTM, neither the trait score estimates nor their corresponding standard errors depend on population distributions. In IRT, trait scores are estimated separately for each score or response pattern, controlling for the characteristics (e.g. difficulty) of the items that are administered. Standard errors are the smallest when items are optimally appropriate for a particular trait score level and when item discriminations are high.

As far as the **test (measurement) length and reliability** are now concerned, in IRT short tests can be more reliable than it is the case for CTM. In CTM (based for example on the Spearman-Brown prophesy formula) a test is lengthened by factor of n parallel parts, and hence the true variance increases more rapidly than error variance. Thus, in CTM, shorter tests (measurements) generally imply increased measurement error. The new rule from IRT asserts that short tests can be more reliable than longer tests. The implication, of course, is that the shorter test yields less measurement error.

Another aspect relates to **interchangeable test forms** when respondents receive different instruments – measurement forms, where some type of equating is needed before their scores can be compared. Traditionally, CTM relied on several conditions associated with the measurement of the form parallelism in order to equate scores. These conditions included the *equality of means* and *variances across measured items,* as well as *equality of covariances with external variables.* In practice, measurement form

parallelism cannot be met. Measurement form either means or variances often differ somewhat. Furthermore, often score comparisons between rather different measurements sometimes are desired. Thus substantial effort must be devoted to procedures for test (e.g. process of scale measurement) equating.

More recent extensions of CTM have considered the test form equating issue more liberally, as score equivalencies between forms. Several procedures have been developed equating tests with different item properties, such as linear equating and equipercentile equating. These methods are used in conjunction with various empirical designs such as random groups or common anchor items.

The IRT version of "equating" follows directly from the IRT model, which implicitly controls for item differences between test forms. Finally, most important, better estimation of trait levels for all individuals are obtained from administering *different* test forms. More accurate estimation of each individual means that score differences are more reliable. Thus, the new rule from IRT means that *nonparallel* test forms (that differ substantially, and deliberately, in difficulty level from other forms), yield better score comparisons.

According to the next condition – **unbiased estimation of item profiles,** in CTM, statistic for item difficulty is typically *p*-value, which is computed as the proportion passing. The CTM statistic for item discrimination is item-total correlation (e.g. biserial correlation). Both statistics can differ substantially across samples if computed from *unrepresentative samples.* In contrast, IRT the correspondence of item difficulty values is quite close between any of the two groups, that is, unbiased estimates of item properties may be obtained from nonrepresentative samples.

Differences appear also when **establishing meaningful scale scores.** It can be noted that test score meaning depends on specifying an appropriate comparison. A comparison is defined by two features:

- the standard with which a score is compared,
- the numerical basis of the comparison (order, difference, ratio, etc.).

For instance in CTM, score meaning is determined by a *norm-referenced standard*, and the numerical basis is order. That is, scores have meaning when they are compared with a relevant group of people for relative position. To facilitate this comparison, raw scores are linearly transformed into standard scores that have more direct meaning for a relative position. An objection that is often raised to norm-referenced meaning is that scores have no meaning for what the person actually can do. In IRT, a score is compared with items, e.g. persons and items are calibrated on a common scale. The meaning of a score can be referenced directly to the items. If these items are

further structured by content, substantive trait level meaning can be derived (Mokken, 1971).

Differences between CTM and IRT appear also in a process of **establishing scale properties.** Routine test development procedures for many social research problems measurements include selecting items to yield *normal distributions* in a target population. Even if normal distributions are not achieved in the original raw score metric, scores may be transformed or normalized to yield a normal distribution. These transformations are nonlinear, and therefore they change the relative distances between scores. And *score distributions* have implications for the level of measurement that is achieved. Jones (1971) pointed out that the classical methods to develop normally distributed trait scale scores will achieve interval scale measurement under certain assumptions. Specifically, these assumptions are that true scores:

- have interval scale properties,
- are normally distributed in the population.

Table 2
Rules of the measurement in Classical Theory of Measurement and Item Response Theory

Classical Theory of Measurement
• The standard error of measurement applies to all scores in a particular population.
• Longer tests are more reliable than shorter tests.
• Comparing test scores across multiple forms depends on test parallelism or adequate equating.
• Unbiased assessment of item properties depends on representative samples from the population.
• Meaningful scale scores are obtained by comparisons of position in a score distribution.
• Interval scale properties are achieved by selecting items that yield normal raw score distributions.
Item Response Theory
• The standard error of measurement differs across scores, but generalizes across populations.
• Shorter tests can be more reliable than longer tests.
• Comparing scores from multiple forms is optimal when test difficulty levels vary across persons.
• Unbiased estimates of item properties may be obtained from unrepresentative samples.
• Meaningful scale scores are obtained by comparisons of distances from various items.
• Interval scale properties are achieved by justifiable measurement models, not score distributions.

Source: Embretson and Reise, 2000

Only linear transformations preserve score intervals as well as distribution shapes (Davison and Sharma, 1990). Thus, if raw scores are normally distributed, then only a linear transformation, such as a standard score conversion, will preserve score intervals to appropriately estimate true score. However, *scale properties* are tied to a specific population. If the measurement is applied to a person from another population, can the interval scale properties still be justified? If not, then scale properties are population-specific.

For IRT models, particularly the Rasch model, interval or even ratio scale properties are achieved to some other extent. The Rasch model also has been linked to fundamental measurement because of the simple additivity of the parameters. A basic tenant of fundamental measurement is additive decomposition (Michel, 1990), in which two parameters are additively related to a third variable. In the Rasch model, additive decomposition is achieved; the log odds that a person endorses or solves an item is the simple difference between his or her trait level, θ_j, and the item's difficulty, b_j will be as follows:

$$\text{LogOdd}_{ij} = \theta_j - b_i. \tag{8.0}$$

In additive decomposition, interval scale properties hold if the law of large numbers apply. Specifically, the same performance differences must be observed when trait scores have the same interscore distances, regardless of their overall positions on the trait score continuum.

8. CONCLUSIONS AND POSSIBLE APPLICATIONS OF THE MEASUREMENT AND SCALES EXTRACTION IN MARKETING FIELD

In both theory and practice, reliability issues are of great importance. A perfectly reliable measurement and thus yielded scale should reflect the researcher's early measurement intentions that remain permanent or are fixed over time. Thus the same measure being applied to the same person or object should give the same value each time, provided that the measured object itself has not changed at this time. However numbers ascribed to units or observations, do not guarantee that the meaning and measurement level will be the same in terms of measured objects. Virtually there is no simple way to look at the numbers and say whether they express any real value or not, or whether there are off the top of the head. This is because some

measurements are actually close to the designated conditions of accidental measurement. Others are not strictly "accidental", but they contain a large element of randomness. The conclusions to be drawn from the same statistical results may vary substantially, depending on whether we know that the measurements are highly accurate or not. Differences and correlation coefficients can often prove to be irrelevant, because measurements were not accurate. So the issues of measurement and reliability deserves much attention from any researcher who cares and endeavors to solve scientific problems.

In literature we can differentiate two main approaches to measurement. One is called The Classical True-Score Theory, the other, Item Response Theory. Appropriate choice of approach depends on the research requirements and objectives (in particular field of science). It also varies from perspective of: 1) type of the measurement, meaning and purpose of applied statistics, 2) unidimensional *vs.* multidimensional measurement, homogeneous (with a single factor measuring only one aspect e.g. human trait) vs. heterogeneous data (dealing with more complex factorial structure e.g. human traits).

As can be easily inferred, each area of science develops its own set of measurement procedures, tools and methods. For example in physics, there are developed methods for detecting subatomic particles. In social sciences, measurement methods are mostly designed and applied to study human life and in general their characteristics. In many social fields (such as business administration, economics, political science, sociology, international relations, communication, etc.), the core aspect of investigations (e.g. based on human activities, behaviour, etc.) involves psychology which indirectly affects all these fields. For example in marketing research, one monitors workers or customers beliefs, motivational states, value systems, expectancies, needs, emotions, or perceptions. In consequence, marketing heavily relies on **psychometrics**.

Psychometrics (as the subspecialty concerned with measuring psychological and social phenomena) has emerged as a methodological paradigm in its own right. Its growth and development was mainly due to:

- widespread use of psychometric definitions of reliability,
- popularity of factor analysis in the social science research,
- adoption of psychometric methods for developing scales measuring an array of various subjects.

That is why psychometric measurement was easily adapted in customers research. Up to this moment it is one of the most common applied solution in

customers measurement and in customers scales based development. This assertion can be proved by Bearden and Netemeyer's compilation work (1999) which classified different types of multi-item scales, developed measures, evaluation procedures, reliability estimation approaches which were used in customer behaviour studies. In many instances topics such as dimensionality, reliability and even validity were broadly discussed according to practical examples and theory that was derived from psychometrics. Bearden and Netemeyer mentioned for example marketing scales related to:

- **customers traits and individual differences,** e.g. 1) Scales related to interpersonal orientation, needs/preferences, and self-concept; 2) Scales related to customer compulsiveness and impulsiveness; 3) Scales related to country image and affiliation; 4) Scales related to customer opinion leadership and opinion seeking; 5) Scales related to innovativeness; 6) Scales related to customer social influence.
- **customers values,** e.g. 1). Scales exploring general values; 2) Scales related to environmentalism and socially responsible consumption; 3) Scales related to values on materialism and possessions/objects.
- **customers involvement, information processing, and price perceptions,** e.g., 1) Scales on involvement with a specific class of product; 2) Scales on involvement general to several products; 3) Scales related to purchasing involvement.
- **customers reactions to advertising stimuli,** e.g. 1) Scales related to ad emotions and ad content; 2) Scales related to ad believability/credibility; 3) Scales related to children's advertising.
- **customers attitudes about satisfaction and post-purchase behavior,** e.g. 1) Scales measuring customers attitudes toward business practices and marketing; 2) Scales related to post-purchase behaviour/discontent; 3) Scales toward product/services satisfaction.

Also beyond the above customers based scales, the other inherited subjects by marketing with its applications (derived from psychometrics) can be found in:

- **job satisfaction measures scales,**
- **role perceptions/conflict in organization scales,**
- **job burnout/tension scales,**
- **performance measures scales,**
- **control and leadership scales,**
- **organizational commitment,** or **sales/selling approaches in the company scales.**

In consequence psychometrics provided solid basis for many research issues in marketing studies. This is because the principles of psychometrics are set in regards to the people and people make up the main area of any marketing researchers scientific inspiration.

Marketing, however, reached its own concept of measurement and scale construction. For instance, at recent times we were supplemented with Rossiter's C-OAR-SE approach (as a sort of direct response to psychometrics) on how to measure marketing phenomenon and construct an appropriate scale. In C-OAR-SE, Rossiter (2002) optionally proposed a brand new procedure for development of the scale to measure marketing latent constructs. He gave a new perspective on the measurement and indicated also when to use single-item vs. multiple-item scales and when to use an index of essential items rather than selecting unidimensional items.

REFERENCES

Allen, M. J., Yen, W. M., *Introduction to Measurement Theory.* Waveland Press, Illinois 1979.

Andrich, D., *A Rating Formulation for Ordered Response Categories,* "Psychometrika", Vol. 43., pp. 561-573, 1978.

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", Vol., 9, pp. 95-104, 1982.

Aranowska, E., *Pomiar ilościowy w psychologii* [*Quantitative Measurement in Psychology*]. Scholar, Warszawa 2005.

Armor, D. J., *Theta Reliability and Factor Scaling* [in:] Costner H. L., (ed.), *Sociological Methodology*. Jossey-Bass, San Francisco 1974.

Bearden, W. O., Netemeyer, R. G., *Handbook of Marketing Scales – Multi-item Measures for Marketing and Consumer Behavior Research.* Sage Publication, London 1999.

Birnbaum, A., *Some Latent Trait Models and their Uses in Inferring an Examinee's Ability* [in:] Lord, F. M., Novick, M. R. (eds.), *Statistical Theories of Mental Test Scores*, pp. 397-479. Addison-Wesley, Reading, MA 1968.

Blalock, H. M., *The Measurement Problem,* [in:] Blalock H. M., Blalock A., (eds.) *Methodology in Social Research.* McGraw-Hill, New York 1968.

Bollen, K. A., *Structural Equations with Latent Variables.* Wiley and Sons, New York 1989.

Brzezinski, J., *Metodologia badań psychologicznych* [*Methodology of Psychological Research*]. PWN, Warszawa 2007.

Cortina, J. M., *What Is Coeffi cient Alpha? An Examination of Theory and Applications.* "Journal of Applied Psychology", Vol. 78, pp. 98-104, 1993.

Cronbach, L. J., *Test Reliability: Its Meaning and Determination,* "Psychometrika", Vol. 12, pp. 1-16, 1947.

Cronbach, L. J., *Coefficient Alpha and the Internal Structure of Tests*, "Psychometrika", Vol. 16, pp. 297-334, 1951.

Davison, M. L., Sharma, A. R., *Parametric Statistics and Levels of Measurement: Factorial Designs and Multiple Regressions,* "Psychological Bulletin", Vol. 107, pp. 394-400, 1990.

DeVellis, R. F., *Scale Development – Theory and Applications*. Sage Publications, London 2003.

Duncan, O. D., *Notes On Social Measurement – Historical and Critical*. Russell Sage, New York 1984.

Embretson, S. E., *A General Latent Trait Model for Response Processes,* "Psychometrika", Vol. 49. pp. 175-186, 1984.

Embretson, S. E., *A Multidimensional Latent Trait Model for Measuring Learning and Change,* "Psychometrika", Vol. 56, pp. 495-516, 1991.

Embretson, S. E., Reise, S. P. *Item Response Theory for Psychologists.* Lawrence Erlbaum Associates, New Jersey 2000.

Ferguson, G. A, Takane, Y., *Analiza statystyczna w psychologii i pedagogice* [*Statistical Analysis in Psychology and Pedagogics*]. PWN, Warszawa 2009.

Graham, J. M., *Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability,* "Psychological Measurement"**,** Vol. 66, 6, pp. 930-944, 2006.

Guilford, J. P., *Psychometric Methods,*McGraw-Hill, New York 1936.

Gulliksen, H., *Theory of Mental Tests.* Wiley, New York 1950.

Hair, J. F., Anderson, R. E., Tatham, R. L., Black W. C., *Multivariate Data Analysis with Readings,* 3rd ed. Macmillan, New York 1992.

Hattie, J.*, Methodology Review: Assessing Unidimensionality of Tests and Items,* "Applied Psychological Measurement"*,* Vol. 9. pp. 139-164, 1985.

Iacobucci, D., Coughlan, A. T., Duhachek, A., *Results on the Standard Error of the Coefficient Alpha Index of Reliability*, "Marketing Science", Vol. 24, No. 2, pp. 294-301, Spring 2005.

Jones, L. V., *The Nature of Measurement* [in:] Thorndike R. L., (ed.) *Educational Measurement,* 2nd ed., American Council on Education, Washington D.C.1971.

Kelderman, H., Rijkes, C. P. M., *Loglinear Multidimensional IRT Models for Polytomously Scored Items,* "Psychometrika", Vol. 59, pp. 149-176, 1994.

Kaydos, W. J., *Operational Performance Measurement*. CRC Press, Florida 1999.

Kelly, G. A., *The Theory and Technique of Assessment*, "Annual Review of Psychology", Volume 9, pp. 323-352, February 1958.

Kozyra, C. *Metody analizy i oceny jakości usług* [*Methods of Analysis and Evaluation of Service Quality*] – PhD Thesis, Akademia Ekonomiczna we Wrocławiu, 2004

Kuder, G., Richardson, M., *The Theory of the Estimation of Test Reliability, "*Psychometrika" pp. 151-160, Vol. II, September 1937.

Lord, F. N., Novick, M. R., *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, MA 1968.

Magnusson, D., *Wprowadzenie do teorii testów* [*Introduction to Test Theory*]. PWN, Warszawa 1981.

Malhotra, N. K., *Basic Marketing Research – A Decision-Making Approach,* 3rd ed. Pearson, London 2009.

Masters, G. N., *A Rasch Model for Partial Credit Scoring,* "Psychometrika", Vol. 47, pp. 149-174, 1982.

Michell, J., *An Introduction to the Logic of Psychological Measurement.* Lawrence Erlbaum Associates, Hillsdale, NJ 1990.

Mokken, R. J. *A Theory and Procedure of Scale Analysis.* De Gruyter, Berlin 1971.

Netemeyer, R. G., Bearden, W. O., Sharma, S., *Scaling Procedures – Issues and Applications.* Sage Publications, London 2003.

Nunnally, J. C., *Psychometric Theory.* McGraw-Hill, New York 1978.

Ostasiewicz, W., *Statistical Modeling of Survey Data.* Technical Reports of Department of Statistics and Economic Cybernetics, No. 37, Wrocław 2002.

Ostasiewicz, W., *Istota pomiaru statystycznego* [*Essence of Statistical Measurement*], [in:] Ostasiewicz W., (ed.) *Pomiar statystyczny.* Akademia Ekonomiczna, Wrocław 2003.

Peter, J. P., *Reliability: A Review of Psychometric Basics and Recent Marketing Practices,* "Journal of Marketing Research", Vol. 16, pp. 6-17, February 1979.

Raju, N. S. *The Area Between Two Item Characteristic Curves, "*Psychometrika", Vol. 53, pp. 495-502, 1988.

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests.* University of Chicago Press, Chicago 1960.

Rosenbaum, P. R. *Comparing Item Characteristic Curves,* "Psychometrika", Vol. 52, pp. 217-233, 1987.

Rossiter, J. R., *The C-OAR-SE Procedure for Scale Development in Marketing,* "International Journal of Research in Marketing", Volume 19, Issue 4, pp. 305-335, December 2002.

Sagan, A, *Zastosowanie wielowymiarowych skal czynnikowych i skal Rascha w badaniach marketingowych* [*Application of Multidimensional Factor Scales and Rasch Scales in Marketing Research*], Zeszyty naukowe, No. 605, pp. 73-92, 2002.

Spearman, C., *The Proof and Measurement of Association Between Two Things*, "American Journal of Psychology", Vol. 15, pp. 72-101, 1904a.

Spearman, C., *General Intelligence Objectively Determined and Measured.* "American Journal of Psychology", pp. Vol. 15, 201-293, 1904b.

Spearman, C., *Demonstration of Formulae for True Measurement of Correlation,* "American Journal of Psychology", Vol. 18, pp. 161-169, 1907.

Spector, P. E., *Summated Rating Scale Construction,* Sage Publications, London 1992.

Symonds. P. M., *Factors Influencing Test Reliability,* "Journal of Educational Psychology", Vol. 19, pp. 73-87, February 1928.

Stevens, S. S., *Mathematics, Measurement and Psychophysics*, [in:] Stevens, S. S. (ed.) *Measurement in Social Sciences: Theories and Strategies.* John and Wiley, New York 1951.

Torgerson, W. S. *Theory and Methods of Scaling.* John Wiley and Sons, New York 1958.

Wertz, C., Linn, R., Jöreskog, K., *Intraclass Reliability Estimates: Testing Structural Assumptions,* "Educational and Psychological Measurement", 34, 1, pp. 25-33, 1974.

Wilson, M., *Constructing Measures: An Item Response Modeling Approach.* New York: Lawrence Erlbaum Associates, 2005.

Zeller, R. A., Carmines, E. G., *Reliability and Validity Assessment.* Sage Publications, New York 1979.

Rost J., *Rasch Models in Latent Classes: An Integration of Two Approaches to Iitem Analysis,* "Applied Psychological Measurement" Vol. 14., pp. 271-282. 1990.

Received: February 2011, revised: April 2011