Journal of Education; ISSN 2298-0172

Two approaches to psychometric process: Classical test theory and item response theory

Mehtap ERGUVEN*

Abstract

In the development of Measurement Theory, there are two main statistical approaches to describe characteristics of an individual and to analyze abilities and latent attributes of a subject which are Classical Test Theory (CTT) and Item Response Theory (IRT). This study provides information about the essential properties of both theories, determines psychometric process of measurement with details, compares models of theories and obviously expresses the advantages and disadvantages of both theories. The earliest theory of measurement, CTT, and the enhanced application of this theory, IRT models are examined from the common and different points of view. This article emphasizes the importance of the constructing, measuring, evaluating and correctly interpreting the educational measurement process.

Keywords: Measurement theory, classical test theory, item response theory, psychometrics, educational measurement

Introduction

Various characteristics of a person are probed and measured periodically through various educational, psychological and measurement tools, including early childhood developmental tests, various aptitude and achievement tests, intelligence tests, behavioral rating scales, etc. (Suen, 1990).

Because the roles of the exemplified tests are so important in the social life, constructing, designing and evaluating educational and psychological tests becomes essential. A good test model might specify the precise relationships among test items and ability scores so that careful test design work can be done to produce the desired test score distributions and errors of the size that can be tolerated (Hambleton and Jones, 1993).

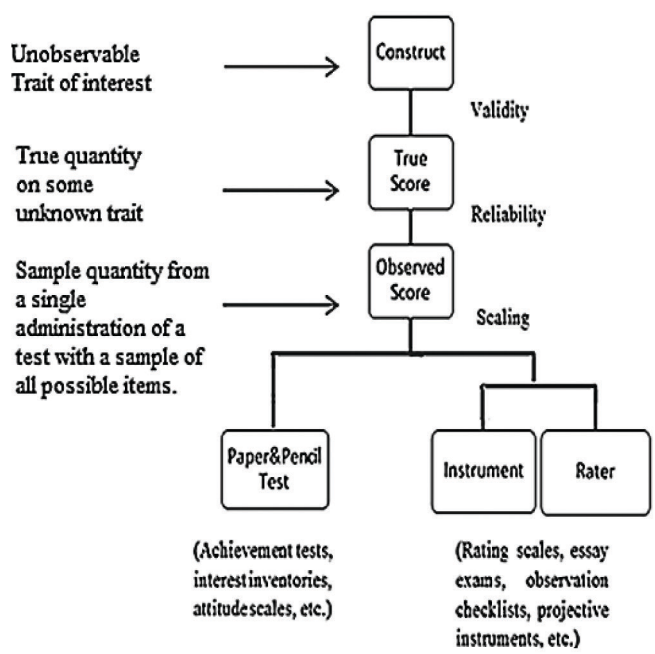

Figure 1
The psychometric process, (Suen, 1990, p.6).

Psychometricians are concerned with the design and development of the tests, the procedures of testing, instruments for measuring data, and the methodology to understand and evaluate the results. The first psychometric instruments were designed to measure the concept of intelligence. Samejima (1997) defines the main objective of psychometrics as mathematical modeling of human behavior.

Identifying cognitive abilities of a test-taker and representing them as a reliable numerical score is the main purpose of educational and psychological measurement. This score is accessible by means of psychometric process. The first step is constructing the exam questions, determining the observed score from that examination, using this observed score to obtain a true score is the third step in this process (**Figure 1**).

If the reliability is high enough, the observed score can be considered as deputy of the true score. Consistency of measurement depends on the reliability of constructed examination. A reliable test, across various conditions and situations, including different evaluators and testing environments, gives approximately the same results.

Validity describes how well one can legitimately trust the results of a test as interpreted for a specific purpose. Despite the enhancements in technology and measurement theory, the requirements for validation have not changed and validation is not optional in measurement.

According to APA standards (AERA, APA, & NCME, 1999, p. 9), “validity refers to the degree to

* Ph.D. Student, Faculty of Computer Technologies and Engineering, International Black Sea University, Tbilisi, Georgia
E-mail: merguven@ibsu.edu.ge.

which evidence and theory support the interpretations of test scores entailed by purposed use of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests".

The term construct refers to the concept, attribute, or variable that is the target of measurement as shown in figure1 at the last step of psychometric process. Most targets of a measurement in psychological assessment, regardless of their level of specificity, are constructs in that they are theoretically defined attributes or dimensions of people.

Construct can be determined as instrument's (examination's) intended purpose, the process for developing and selecting items (the individual questions, prompts, or cases comprising the instrument) or the wording of individual items and qualifications of item writers and reviewers (Haynes, Richard, and Kubany, 1995). Validity is a property of inferences, not instruments; therefore validity must be established for each intended interpretation (Cook, D.A.; Beckman,T.J., 2006). Unfortunately, instrument's scores reflect the underlying construct sometimes accurately or less accurately but never perfectly.

The purpose of a measurement is representing individual's properties using valid and adequate theoretical models with respect to reliability and after administration of a test, interpreting the obtained outputs in a scientific manner. One of the most striking and challenging phenomena in the Social Sciences is the unreliability of its measurements: Measuring the same attribute twice often yields two different results (Steyer, 1999). Science is based on the adequacy of its measurement. Poor measures provide a weak foundation for research (Foster and Cone, 1995). A basic distinction in science may be made between theoretical constructs and observed measures thought to represent these constructs (Revelle, 2013). Often the target of the measurement and the result of this process do not fit each other.

The accuracy of the obtained information is related to the capacity and modernity of the applied techniques. A relationship between theory which is used for construction and implementation of a test, and measurement and evaluation of a test is critical to provide realistic and adequate information about a subject.

A test can be studied from different angles and the items in the test can be evaluated according to different theories. There are several theories to analyze and manipulate in whole psychometric process. Two such theories will be discussed in this study. The main characteristics of these theories, relations between them, and basic properties of theories with their advantages and disadvantages and differences among them are discussed and represented in the next sections of the article.

Classical Test Theory

According to Bejar (1983), random sampling theory and item response theory are two major psychometric theories for the study of measurement procedures. In random sampling theory, there are two approaches, the classical theory approach and the generalizability theory approach. A CTT (also known as classical true score theory) is a simple model that describes how errors of measurement can influence observed scores (Marcoulides, 1999).

Classical test theory (Gulliksen, 1950) is the earliest theory of measurement. For a long time psychometric characteristics of personality measures have been examined using CTT assumptions. The major target of this theory is estimating the reliability of the observed scores of a test. If the test is applied on a particular sample of items, at that particular time, in the reliable conditions, this exam gives an observed score of the examinee. Under all possible conditions at various times, using all possible similar items, the mean of all these observed scores would be the most unbiased estimate of the subject's ability. Thus, mean is defined as the true score. In any single administration of a test, the observed score is most likely different from the true score (Suen, 1990). This difference is called random error score. In the framework of CTT each measurement (test score) is considered being a value of a random variable X consisting of two components: a true score and an error score (Steyer, 1999). This relationship is represented in below formula:

$$X = T + E$$

Because the true score is not easily observable, instead, the true score must be estimated from the individual's responses on a set of test items.

In CTT, the observed score is assumed to be measured with error. However, in developing measures, the goal of CTT is to minimize this error (McBridge, 2001) . In that case, importance of a reliability of a test and calculating the reliability coefficient increases. if we know reliability coefficient, we can estimate the error variance. The square root of error variance is determined as standard error of measurement (SEM) and helps to define the confidence interval to have a more realistic estimation of the true score.

Reliability is considered an attribute of the test data and not the assessment itself in CTT. In fact, APA standards (AERA, APA, & NCME, 1999) state that when reliability is reported, it must be accompanied by a description of the methods used to calculate the coefficient, the nature of the sample used in the calculations, and conditions under which the data were collected. However, reliability estimates calculated through these procedures are sample dependent and, as a result, have a number of practical limitations when building or evaluating technology-enhanced assessments (Scott and Mead, 2011).

The alpha formula is one of several analyses that may be used to gauge the reliability (i.e., accuracy) of psychological and educational measurements. This formula was designed to be applied to a two-way table of data where rows represent persons (p) and columns represent scores assigned to the person under two or more conditions (i). Because the analysis ex-

amines the consistency of scores from one condition to another, procedures like alpha are known as internal consistency analyses (Cronbach and Shavelson, 2004). The reliability was computed with coefficient alpha, defined as:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_X^2}\right),$$

where n is number of items in the test, σ_i^2 is the variance on item i and σ_x^2 is the variance on the overall test result (Wiberg, 2004).

Cronbach's α can be shown to provide a lower bound for reliability under rather mild assumptions. Thus, the reliability of test scores in a population is always higher than the value of Cronbach's α in that population. A value of 0.7-0.8 is an acceptable value for Cronbach's α; values substantially lower indicate an unreliable scale.

There are two indices in CTT, ***"p"*** and ***"r"***. The proportion of examinees passing an item is called difficulty index **p**, actually, high values of **p** indicates an easy item. The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as item discrimination ***"r"***, which is often expressed statistically as the Pearson product-moment correlation coefficient between the *scores on the item* (e.g., 0 and 1 on an item scored right-wrong) and the scores on the total test. If an item is dichotomously scored, this estimate is often computed as a *point-biserial correlation coefficient* (Fan, 1998). The formula of point biserial correlation (r_{pbi}) is defined by:

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{p}q$$

where, Mp = whole-test mean for students answering item correctly (i.e., those coded as 1s),
Mq = whole-test mean for students answering item incorrectly (i.e., those coded as 0s),
St = standard deviation for whole test,
p = proportion of students answering correctly (i.e., those coded as 1s),
q = proportion of students answering incorrectly (i.e., those coded as 0s) (Brown, 2001).

Point biserial correlation (r_pbi) ranges from -1 to +1. A high point-biserial coefficient means that students selecting the correct response are students with higher total scores, and students selecting incorrect responses to an item are associated with lower total scores. According to the value of r_pbi, item can discriminate between low-ability and high-ability examinees. Very low or negative point-biserial coefficients help to identify defective items.

Item Response Theory and its Models

Item Response Theory (IRT) is used in a number of disciplines including sociology, political science, psychology, human development, business, and communications, as well as in education where it began as method for the analysis of educational tests (Templin, 2012).

CTT was originally the leading framework for analyzing and developing standardized tests. Since the beginning of the 1970's IRT has more or less replaced the role CTT had and is now the major theoretical framework used in this scientific field (Wiberg, 2004).

IRT allows the user to specify a mathematical function to model the relationship between a latent trait, θ, and the probability that an examinee with a given θ will correctly answer a test item. Until the 1980s, IRT research focused largely on the estimation of model parameters, the assessment of model-data fit, and the application of these models to a range of testing problems using dichotomously scored (yes/no,1or 0) multiple-choice items. Research on performance assessments, polytomous response formats, and multidimensional traits began in earnest, as did work on computerized adaptive testing. An outcome of this expanded focus was a host of new IRT models that allowed researchers to tackle complex problems, not only in achievement testing, but also in areas such as attitude, personality, cognitive, and developmental assessment (Gierl and Bisanz, 2001).

The first consideration when choosing the right model involves the number of item response categories. For dichotomous items, the 1, 2, and 3 parameter logistic models are most common (1PL, 2PL, 3PL), and models including an upper asymptote parameter (e.g., 4PL) are also possible. For polytomous items, variations of the Partial Credit Model, Rating Scale Model, Generalized Partial Credit Model, and Graded Response Model are available for ordered responses, and the Nominal Model is appropriate for items with a non-specified response order.

The second important consideration when choosing the right model is whether the item discrimination parameters, or slopes, should be free to vary across items, or whether a model from the Rasch (Rasch, 1960) family is more appropriate (Edelen and Reeve, 2007). The IRT model (1PL, 2PL, 3PL) can be defined using the 3PL model formula:

$$\boldsymbol{P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \qquad i = 1, 2, \ldots, n,}$$

where $\boldsymbol{P_i(\theta)}$ is the probability that a given test-taker with ability $\boldsymbol{\theta}$ answer a random item correctly, a_i is the

item discrimination, bi is the item difficulty and ci is the pseudo guessing parameter (Hambleton, Swaminathan, and Rogers, 1991). The 2PL model is obtained when c = 0. The 1PL model is obtained if c = 0 and a = 1 (Wiberg, 2004) or constant.

Items should be selected at any point in the testing process to provide maximum information about an examinee's ability. In this application, a model is needed that places persons and items on a common scale (this is done with item response theory models) (Hambleton and Jones, 1993). In IRT, higher levels of information are produced when items have higher discrimination "a" parameters, and smaller lower-asymptote "c" parameters (Harvey & Hammer, 1999).

A "b" parameter defines how easy or how difficult is an item and an "a" parameter determines how effectively this item can discriminate between highly proficient students and less-proficient students. The guessing parameter "c" determines how likely the examinees are to obtain the correct answer by guessing (Yu, 2013).

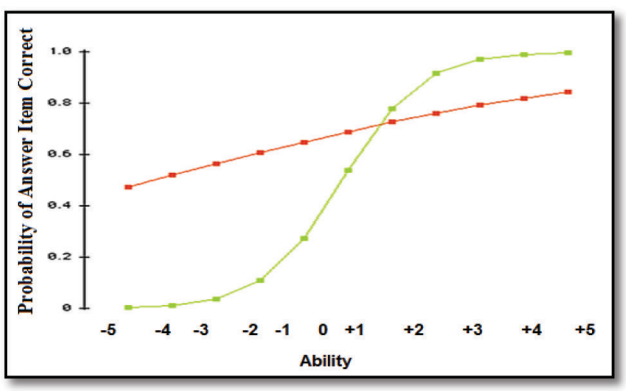

Figure 2: *ICCs of low and high discrimination (Low discrimination in the red curve, high discrimination in the green curve).*

IRT Assumptions

Before using IRT models in psychometric process, two basic assumptions must be met. These are unidimensionality and local independency. The assumption of unidimensionality means that only one trait or ability is measured by the items. Local independency and unidimensionality are similar, but not equivalent, concepts. When the assumption of unidimensionality is met, so is the assumption of local independence. However, the assumption of local independence can be met without unidimensional data as long as all aspects that affect the test results are taken into account (McBridge, 2001).

In the local independency assumption, responses for different items are not related. An item does not provide any clue to answer another item correctly. If local dependence does exist, a large correlation between two or more items can essentially affect the latent trait and it causes lack of validity.

In the following two figures, relations between the examinee with θ ability, (E_θ) and his/her responses to different items (i1, i2, i3, i4) are represented within two situations: dependent and independent.

In figure 3, in response to a question, the examinee has no chance to reply any item with the help of another item. Items are independent and they do not contain any hint among each other.

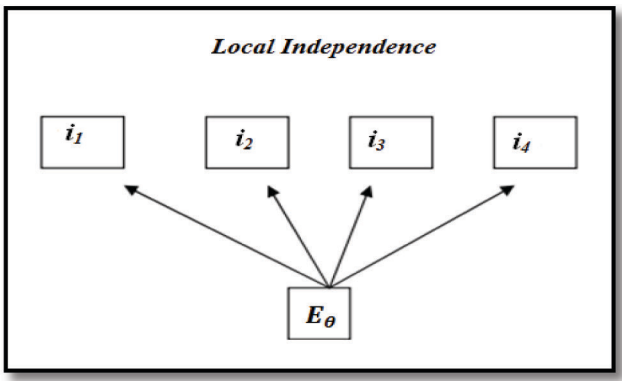

Figure3: *Illustration of independent items and subject.*

In figure 4, test-taker answers item 1 and item 4 with his/her knowledge and ability, but item1 contains information to solve question 3 and item 4 gives clues to answer item 2. Therefore, such questions should be eliminated, since they violate the local independence assumption of IRT. Such questions are not adequate to estimate an examinee's ability accurately.

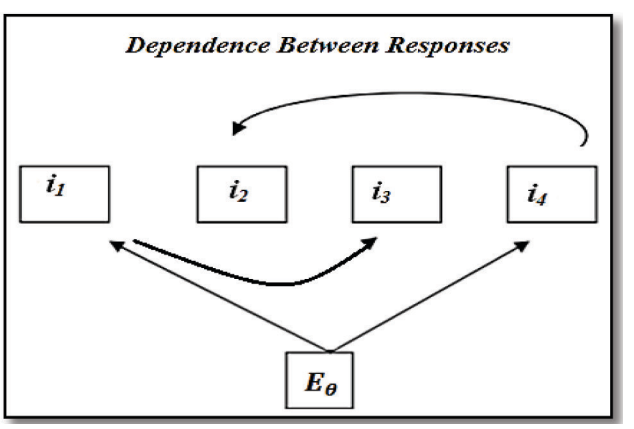

Figure 4: *Illustration of dependent items and examinee*

The third assumption of IRT is Item Characteristic Curve (ICC). The monotonically increasing item characteristic function specifies that the examinees with higher scores on the traits have higher expected probabilities for answering the item correctly than the examinees with lower scores on the traits. In the one-trait or one-dimensional model, the item characteristic function is called item characteristic curve (ICC) and it provides the probability of examinees answering an item correctly for examinees at different points on the ability scale. In addition it is common to assume that, ICCs are described by one, two, or three parameters (Hambleton, 1982) .

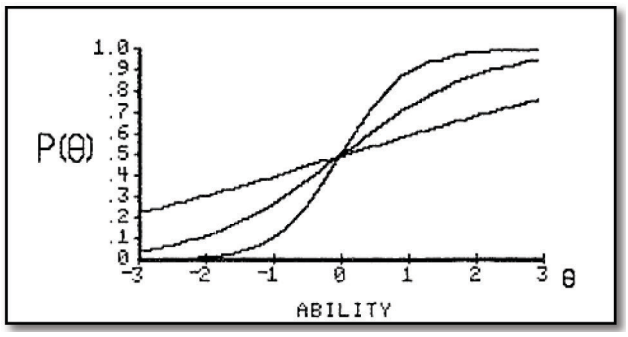

Figure 5: *Three item characteristic curves, with the same difficulty but different discrimination parameters.*

Comparison of CTT and IRT Models, Advantages and Disadvantages:

Benefits obtainable through the application of classical test models to measurement problems include:

1. Smaller sample sizes required for analyses (a particularly valuable advantage for field testing),
2. Simpler mathematical analyses compared to item response theory,
3. Model parameter estimation, which is conceptually straightforward, and
4. Analyses, which do not require strict goodness-of-fit studies to ensure a good fit of model to the test data (Hambleton and Jones, 1993).

Beside these properties, CTT has several limitations.

• The most challenging critique of many applications of CTT is that they are based on rather arbitrarily defined test score variables. If these test score variables are not well chosen, any model based on them is not well-founded, either. In most applications, the decision how to define the test score variables "Yi" on which models of CTT are built is arbitrary to some degree. It should be noted, however, that arbitrariness in the choice of the test score variables cannot be avoided altogether. Even if models are based on the item level, such as in IRT models, one may ask "Why these items and not others"? Whether or not a good choice has been made, will only prove in model tests and in validation studies. This is true for models of CTT as well as for models of alternative theories of psychometric tests (Steyer, 1999).

• Another limitation of classical test theory is that scores obtained by CTT applications are entirely test dependent and unfortunately p and r statistics are dependent on the examinee sample from which they are obtained. Among the greatest advantages of the IRT over the CTT are: the possibility of comparing between the latent traits of individuals of different populations when they are submitted to tests or questionnaires that have certain common items; it also allows for the comparison of individuals of the same population submitted to totally different tests; this is possible because the IRT has the items as its central elements, not the tests or the questionnaire as a whole; it allows for a better analysis of each item that makes up the measure (Araujo, Andrade, and Borlotti, 2009).

• IRT models based on an explicit measurement models. A major limitation of traditional assessment frameworks is the assumption that measurement precision is constant across the entire trait range. IRT models, however, explicitly recognize that measurement precision may not be constant for all people (Fraley, Waller, and Brennan, 2000.

• Classical test models are often referred to as "weak models" because the assumptions of these models are fairly easily met by test data. (Though, it must be mentioned that not all models within a classical test theoretic framework are "weak." Models such as the binomial test model, which are based upon a fairly restrictive assumption about the distribution of error scores, are considered "strong models.") Item response models are referred to as strong models too (Hambleton and Jones, 1993).

• IRT and the CTT person parameters are highly comparable and also item difficulties and item discriminations are very comparable. This comparability is defined by Courville (2005), Fan (1998) and MacDonald and Paunonen (2002).

• Accordingly, within the CTT framework, the question of model validity is almost never addressed (Progar and Sočan, 2008).

• The combination of Computerized Adaptive Testing (CAT) and IRT provides several advantages. Item banks contain information on the wording of each item, the concept it measures, and its measurement characteristics according to a measurement model. Most CAT-based assessments utilize a set of statistical models building on IRT to select items and to score the responses. By selecting the most appropriate items for each person, assessment precision are optimized for a given test length and irrelevant items can be avoided.

Assessment precision can be adapted to the needs of the specific application. For example, for a diagnostic purpose precision should be high for scores close to diagnostic cut-points, or test precision could be set high over all the score range for the purposes of follow-up of individuals. At the end of the assessment, the respondent can be given a score immediately, along the guidelines on how to interpret the score (Bjorner, Kosinski and Ware, 2004).

Main Differences between CTT and IRT Models

In this chapter fundamental differences of both theories are described. These basic distinctive attributes are presented in Table 1.

At the item level, the CTT model is relatively simple. CTT does not invoke a complex theoretical model to relate an examinee's ability to success on a particular item. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item (assuming it is dichotomously scored) (Fan, 1998).

Table 1. *Main Difference between CTT and IRT Models, source: (Hambleton, R.K.; Jones, R.W., 1993)*

Area	Classical Test Theory	Item Response Theory
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak(i.e, easy to meet with test data)	Strong (i.e, more diffucult to meet with data)
Item-ability relationship	Not specified	Item characteristics functions
Ability	Test scores or estimated true scores are reported on the test-score scale (or a transformed test score scale)	Ability scores are reported on the scale -∞ to +∞ (or a transformed scale)
Invariance of item and person statistics	No--item and person parameters are sample dependent	Yes—item and person parameters are sample independent, if model fits test data
Item statistics	p, r	b, a, and c (for the three-patameters model) plus corresponding item information functions
Sample size (for item parameters estimation)	200 to 500 (in general)	Depends on te IRT model but larger samples, i.e, over 500, in general, are needed

Lord and Novick (1968) made the important observation that examinee observed scores and true scores are not synonymous with ability scores. The main idea is that examinees come to a test administration with ability levels or scores in relation to the construct being measured by the test. These ability scores (in IRT) are test-independent. However, examinee test scores and corresponding true scores will always depend on the selection assessment tasks from the domain of assessment tasks over which their ability scores are defined. Examinees will have lower true scores on difficult tests and higher true scores on easier tests, but their ability scores remain constant over any tests that might be built to measure the construct (Hambleton and Jones, 1993).

Classical test theory can be defined as "test based," whereas IRT can be defined as "item based".

CTT and its models are not really adequate for modeling answers to individual items in a questionnaire. This purpose is more adequately met by models of item response theory (IRT) which specify how the probability of answering in a specific category of an item depends on the attribute to be measured, i.e., on the value of a latent variable (Steyer, 1999).

An important distinction between IRT and CTT is that IRT defines a scale for the underlying latent variable that is being measured by a set of items, and items are calibrated with respect to this same scale. This is why IRT is said to have a "built-in" linking mechanism (Edelen and Reeve, 2007).

In particular, the focus on estimating an ICC for each item provides an integrative, holistic view of the performance of each item that is not readily available when using CTT-based methods to develop or examine a test.

(CTT) can measure the difficulty level and discrimination power of any item; they have been generally recognized as sample dependent.

IRT models consist of invariance of ability and item parameters. Examinee trait (ability) level estimates do not depend on which items are administered, and in turn, item parameters do not depend on the group of examinees.

Whereas in CTT a single number (e.g., the internal-consistency reliability coefficient, or the SEM based on that reliability) would be used to quantify the measurement-precision of a test, a continuous function is required in IRT to convey comparable data, given that in the IRT approach, a test need not be assumed to possess a constant degree of measurement-precision across the entire possible range of scores (Harvey and Hammer, 1999).

Conclusion

Multiple raters, the psychological state of test- taker, environmental factors or test itself affect examinees' scores in each implementation of instrument. Sometimes, each test administration gives different results about the same person. The only valid and reliable constructions of examinations are for interpreting the real aspect of the ability of individual.

As it has been mentioned before, the main purpose of the psychometric process and usage of different measurement theories is to determine maximum information about an individual. This valuable information is accessible by different methods, if valid, theoretic mathematical background of implementation is used and a reliable atmosphere is satisfied. Both CTT and IRT are scientific methods which have a pioneer role in educational measurement and psychometric process. Essential rules of these theories are dis-

cussed and presented in this study.

CTT has served the measurement community for most of this century and IRT has witnessed an exponential growth in recent decades (Fan, 1998). Therefore, focus of the study is representing the main principles of these theories, and determining their effects on the educational measurement process.

In the comparison of theories it is determined that IRT models are more informative than CTT models if samples are big enough to allow their application, if the items obey the laws defining the models, and if detailed information about the items (and even about the categories of polytomous items, such as in rating scales) is sought (Steyer, 1999).

Besides depicting the simplicity of the CTT model from multiple points of view, various limitations of the model are determined. Fundamental assumptions of CTT and IRT, and differences among them are illustrated. These differences are detailed in item, person and ability level. In the implementation of computerized adaptive testing and questionnaires, adequacy and ascendency of IRT models are underlined.

References

AERA, APA, & NCME (1999). Standards for educational and psychological testing. Washington, D.C.

Araujo,E.P., Andrade, D.F., and Bortolotti, S. (2009, August 24). Item Response Theory. Brazil. 1000-1008

Bejar, I. (1983). Subject matter experts' assessment of item statistics. Applied Psychological Measurement, 7, 303-310

Bjorner, J.B.; Kosinski, M.; Ware, J.E. (2004). Computerized Adaptive Testing and Item Banking

Brown, J. (2001). Statistics Corner: Questions and Answers About Language Testing Statistics: Point--Biserial Correlation Coefficients. Shiken: JLT Testing & Evlution SIG Newsletter, 5(3), 13-17

Cook, D.A.; Beckman,T.J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. The American Journal of Medicine, 119(2), 166.e7-166.e16

Courville, T. G. (2005). An empirical comparison of item response theory and classical test theory item /person statistics. Doctoral dissertation, Texas A&M University. Retrieved November 5, 2007 from http://txspace.tamu.edu/bitstream/handle/1969.1/1064/etd-tamu-2004B-EPSY-Courville-2.pdf?sequence=1.

Cronbach, L.J.; Shavelson, R.J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. Educational and Psychological Measurement, 64(3), 391-418

Edelen, M.O; Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research, 16, 5-18

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educational and Psychological Measurement. June 1998 v58, n3, 357-374

Foster S.L.; Cone J.D. (1995). Validity issues in clinical assessment. Psychol. Psychological Assessment, 248-260

Fraley, R.C.; Waller, N.G; Brennan, K.A.. (2000). An Item Response Theory Analysis of Self-Report Measures of Adult Attachment. Personality and Social Psychology, 78(2), 350-365

Gierl, M. J; Bisanz, J. (2001, December). Item Response Theory for Psychologists-Book Review. Applied Psychological Measurement, 25(4), 405-408

Gulliksen, H. (1950). Theory of mental tests. Wiley, New York

Hambleton, R. (1982). Item Response Theory, The Three Parameter Logistic Model. CSE, National Institute of Education, California

Hambleton, R.K., Swaminathan, H. and Rogers, H.J. (1991). Fundamentals of Items Response Theory. Newbury: Sage

Hambleton, R.K; Jones, R.W. (1993). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Items, 253-262.

Harvey, R., & Hammer, A. (1999). Item Response Theory. The Counseling Psychologist, 27(3), 353-383

Haynes, S.N; Richard, D.C.S; Kubany, E.S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. Psychological Assessment, 7(3), 238-247

Lord, F. M., Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person parameters based on item response theory versus classical

test theory. Educational and Psychological Measurement, 62, 921–943

Marcoulides, G. (1999). Generalizability Theory: Picking up Where the Rasch IRT Model Leaves off? In S. Embretson, & S. Hershberger, The New Rules of Measurement: What Every Psychologist and Educator Should Know. Lawrence Erlbaum Associates Inc., 129-130

McBridge, N. (2001, May). An Item Response Theory Analysis of the Scales from the International Personality Item Pool and the Neo Personality Inventory. Virginia, USA

Progar, Š.; Sočan, G. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. Horizons of Psychology, 17(3), 5-24

Rasch, G. (1960). Probabalistic Models for Some Intelligence and Attainment Tests. Copenhagen Danmarks paedagogiske institut, Denmark

Revelle, W. (2013). Personality Project. Retrieved October 2013, from http://personality-project.org/revelle/syllabi/405.old.syllabus.html

Samejima, F. (1997). Departure from normal assumptions: a promise for future psychometrics with substantive mathematical modeling. Psychometrika, 62, 4,471-493

Scott, J.S.; Mead, A.D. (2011). Foundations for Measurement. In N. Tippins, & S. Adler, Technology-Enhanced Assessment of Talent (1st ed.). CA, USA: Jossey-Bass,Wiley, 21-66

Steyer, R. (1999). Steyer, R. Classical (Psychometric) Test Theory. Jena, Germany

Suen, H. K. (1990). Principles of Test Theories. Hillsdale, NJ: Lawrence Erlbaum

Templin, J. (2012, July 9-13). Item Response Theory. Colorado, Georgia, USA

Wiberg, M. (2004). Classical Test Theory vs.Item Response Theory; An Evaluation of the Test in the Swedish Driving-Lcience Test. Sweden

Yu, C. (2013, August). A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling. Retrieved December 18, 2013 from http://www.creative-wisdom.com