Original Article

Invariance of item difficulty Parameter estimates based on Classical Test Theory and item Response Theory

Rita Rezaee[1], Majid shafiayan[2], Peyman Jafari[3], Nahid Zarifsanaiey[4, 5*]

[1] Clinical education research Center, Health human resources research Center, Shiraz University of medical sciences, Shiraz, Iran. [2] Student Research Committee, Shiraz University of Medical Sciences, Shiraz, Iran. [3] Biostatistics Department & Member of Safety and Health Office, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran. [4] Department of E-learning, Virtual School, Center of Excellence for e-Learning in Medical Sciences, Shiraz University of Medical Sciences, Shiraz, Iran, [5]Virtual university of Medical Science, Tehran, Iran.

Correspondence: Nahid Zarifsanaiey. Department of E-learning, Virtual School, Center of Excellence for e-Learning in Medical Sciences, Shiraz University of Medical Sciences, Shiraz, Iran. Email address:nzarifsanaee@gmail.com

ABSTRACT

Classical Test Theory (CTT) and Item Response Theory (IRT) are two competing theory in measurement field. Superiority of invariance property of item difficulty parameter in item response theory to classical test theory was proved theoretically. This qualitative empirical study is designed to answer how invariant the item difficulty parameter derived from each measurement theory across different samples? Material and Method The present study empirically examined, using norm-referenced large scale data of 39th comprehensive Pre-training examination of medical students of IRAN with 2075 examinee to show invariance property of the item difficulty parameter under the two competing measurement theory. Winestep version 3.71 and SPSS version 17 softwares were used for item analysis in two models. Joint maximum likelihood procedure with Expectation Maximization was applied in this study. Results The findings indicated that the degree of invariance of the item difficulty parameter across different samples, usually considered as the theoretical superiority IRT models, also appeared to be similar for the two measurement frameworks. Discussion The findings suggest that across samples the degree of invariance of the CTT item difficulty index was very similar with IRT item difficulty estimates.

Keywords: Invariance property, Item difficulty, CTT, IRT

Introduction

Invariance of Item Difficulty Parameter estimates based on Classical Test Theory and Item Response Theory. There are two competing measurement theory in the theory of measurement, Classical Test Theory and Item Response Theory. The statistical analysis concerning each framework reflects the differences between two theories. Classical test theory is considered as the true score theory. The assumption that organized effect between answers of examinee is due just to difference in ability of interest is the initial point of the theory. Internal and external condition of examines showing subsets of all other potential sources of differences in the testing materials are assumed either to be constant through robust standardization or to have a disorganized effect or accident by intention [1]. Lord made the remarkable consideration that ability score are not synonymous with examinee observed score and true score. As ability scores are test independent, it is fundamental whereas observed scores and true scores are test dependent [2]. The main concept of the classical test theory is that observed test scores (To) are composed of a true score (T) and an error score (E) were the true and error scores are independent. It was best illustrated in the equation: TO= T+E and the variables are established by spearman (1904) and Novick (1966). It means that examinees with different ability levels or scores pertaining to the test construction participate in a test. Therefore, choice of assessment tools affects examinee test scores and corresponding true scores. While examinees ability scores remain constant over any test might be built to measure the construct, they will have lower true scores on difficult test and higher true scores on easier tests [3]. The standard deviation of errors as the basic

Access this article online	
Website: www.japer.in	**E-ISSN:** 2249-3379

How to cite this article: Rita Rezaee, Majid shafiayan, Peyman Jafari, Nahid Zarifsanaiey. Invariance of Item Difficulty Parameter estimates based on Classical Test Theory and Item Response Theory. J Adv Pharm Edu Res 2018;8(S2):156-161.
Source of Support: Nil, Conflict of Interest: None declared.

measure of error is used in classical test theory. It is usually called the standard error of measurement. To estimate the standard error of measurement, the standard deviation of the observed score and the reliability of the test are applied [4]. According to Adedoyin and Nenty, the true scores reflects what the examinee actually knows, but it is always contaminated by different sources of errors [5]. The test reliability is expressed as a ratio between the true score variance and observed score variance" [5]. In CTT framework, calculating difficultly and discriminative parameters is the cornerstone of item analysis. To estimate the item difficulty of an item, the proportion of examinees endorsing a dichotomous item by choosing the correct response is calculated. The item mean, item difficulty, or item p value is referred to the rate of item endorsement. This value indicates an easy item, approaching 1.0 and indicates a difficult item approaching 0.0. [6]. Evaluation the test is the basic methods of analysis on classical test theory. To indicate item difficulty, frequency of correct responses is computed and the emphasis is on the total test score [7-9]. Under CTT framework, item statistics and item discrimination and test statistics such as test reliability are sample dependent (i.e, dependent on the examinee sample) [10]. According to Fan, The major of CTT can be summarized as circular dependency: (a) The person statistic (i.e., observed score) is (item) sample dependent, and (b) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent. This circular dependency poses some theoretical difficulties in CTT`s application in some measurement situations (e.g., test equating, computerized adaptive testing)" [11]. Because of lack of information related to prediction to examinee performance on a specified item, the CTT framework is unable to designate the test to determine an examinee's proficiency level. And as item parameter indices are sample dependent, it lacks invariance of item parameters across groups of examinees [12]. Simplicity is the main advantage of CTT: to estimate the measured trait, scored responses should be added to items. The true score can be defined operationally as the average score on the infinite number of equivalent repetitions of the measurement, despite the true score is not empirically observable [2]. According to Hambleton, Advantages of many classical test models are that they are based on relatively weak assumptions (i.e., they are easy to meet in real test data) and they are well-known and have a long track record [13]. On the other hand, both person parameters (i.e., true scores) and item parameters (i.e., item difficulty and item discrimination) are dependent on the test and the examinee sample, respectively, and these dependencies can limit the utility of the person and item statistics in practical test development work and complicate any analyses [14].

Item Response theory **(**IRT) has obtained considerable development in recent decades though CTT has served the psychometric community for most of the 20th century [11, 13]. IRT is a model-based paradigm modeling the link between latent trait and item response. In IRT analysis, finding an accurate model rather than precise estimation is a key feature, in this framework, item statistics do not depend on the examinees and person statistics do not depend on the items. So, person parameters are invariant across items, and item parameters are invariant in different sample of examinees [14]. According to fan (1998) IRT is more theory grounded and models the probabilistic distribution of examinees `success at the item level. As its name indicates, IRT primarily focuses on the item level information in contrast to the CTT`s primary focus on test level information. The IRT framework encompasses a group of models, and the applicability of each model in particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items [11]. As CTT emphasizes process of dependability rather than measurement, it is not reliable to item difficulty parameter and calibration nor on total score for estimating the ability, IRT offer a model instead of classical theory [14]. IRT may be considered as latent trait theory sometime. It is referred as the robust true score theory or modern mental test theory because IRT makes stronger assumptions in comparison to classical test theory. Item analysis in this approach considers the probability of getting particular items dichotomously. Each item on a test is based on own item characteristic cure describing the probability of getting each item dichotomously relating to ability of the examinees [3]. IRT is based on two fundamental assumptions. First, an examinee having more ability has greater probability of success getting item right in comparison to less able examinee. Second any examinee always likely to be better on an easier item than on a more different one. In IRT item difficulty is the parameter influencing examinee responses and examinee ability is the feature influencing item difficulty parameter [13]. The basic feature of IRT is that item performance is related to examinee's latent trait [15]. According to Magno (2009) A latent trait is symbolized as the (θ) which refers to a statistical construct. In cognitive test, latent traits are called the ability measured by the test. Ten total score on a test is taken as an estimate of that ability. A person's specified ability (θ) succeeds on an item of specified difficulty" [16].

In IRT framework, the Rash model is appropriate for modeling dichotomous responses and the probability of an examinees` correct response, in this one parameter logistic model, all items discriminations are assumed to be equal to one. The logistic item characteristic curve forms the probability answering an item dichotomously [17]. In the one-parameter IRT model, the only parameter to be estimated is the item difficulty b_i. The one-parameter IRT model formula is

$$Pi\ (\theta) = \frac{e^{Da\ (\theta - bi)}}{1 + e^{Da(\theta - bi)}}$$

Where D is a scaling constant (usually D=1.702). The item discrimination parameter a_i is constrained such that all items have equal and fixed discrimination level a [6]. IRT conquers the main weakness of CTT, the circular dependency of CTT`s item/ person statistics` theoretically. Consequently, in theory, IRT models produce item estimates free of person samples and person statistics free of item samples. This invariance property

of item and person statistics of IRT has been proved theoretically. The item parameter invariance is the main advantage of IRT making distinction between IRT and CTT. The invariance property of IRT item parameters makes it theoretically enable to solve main measurement problems to CTT framework like test equating and computerized adaptive testing [13,18].

The cornerstone of specific objectivity in measurement community is invariance property and lack of it makes many questions about reliability of educational measurement. According to Fan (1998), it is somewhat surprising that empirical studies examining and/or comparing the invariance characteristics of item statistics from the two measurement frameworks are so scarce. It appears that the superiority of IRT over CTT in this regard has been taken for granted in the measurement community, and no empirical scrutiny has been deemed necessary. The empirical silence on this issue seems to be anomaly [11].

There is a lack of empirical knowledge about how, and to what extent, the IRT and CTT item parameter estimates are invariant across different samples of examinees. The purpose of this qualitative empirical comparison is to find out whether the item difficulty parameter estimates are invariant across different samples of person on samples extracted from real data of large scale norm referenced 39th comprehensive pre-training examination of medical students of IRAN using CTT and IRT theoretical framework. To determine whether or not, the item difficulty parameter estimates based on CTT and IRT are significantly invariant across different samples of examinees one research question and two research hypotheses were tested using Pearson correlation at an alpha level of 0.001.

Q: Which of two item difficulty parameter estimates based on CTT or IRT are invariant across different samples of examinees?

H_1: Different samples of examinees have no significant influence on the item difficulty parameter estimates based on CTT across different samples of persons.

H_2: Different samples of examinees have no significant influence on the item difficulty parameter estimates based on IRT across different samples of persons.

Materials and Methods

The data used in this study are from the 39th comprehensive pre-vocational Assessment of medical students in October 2014 to third-grade students in IRAN. This assessment is taken by over 2000 medical student each year. Nearly 50 medical schools of IRAN recommend that applicants submit comprehensive pre-vocational Assessment result.

This paper and pencil four-option multiple choice test is composed of 200 items in 16 sub-tests: General surgery, Gynecology, Pediatric diseases, Internal medicine diseases, Orthopedy, Urology, Ophtomolgy, ENT, Biostatistics & Epidemiology, Pharmacology, Brain and Nerves diseases, Infectious diseases, Radiology, Pathology, Psychiatry, and Skin diseases, distributed into two different form test booklets. The total of 2075 examinees participated in the 39th comprehensive pre-training test was targeted in this study. The details of participant were entered in the table 1.

Table 1. The number of medical students participated in the 39th comprehensive pre-training test.

Major	Female	Percent	male	Percent	Total
Medicine	883	42.5	1192	57.5	2075

Participation in comprehensive pre-training test, subjected to the success in all sectors and subjects tested in the clinical training phase. The passing score of the theoretical courses, clinical in tern ships, and related wards of the third stage of medical education was 12 out of 20. Average total of 14 in the third stage of medical education was required to participate in the exam.

The grade threshold for the exam 70% of the mean of 5% of examinees has achieved the highest score all over the country. Therefore, it was a norm-referenced test. The test was constructed based on syllabuses approved by Postgraduate Medical Planning Council. The examinees must answer test booklet A or B in 200 minutes.

The minimum sample size of examinees proposed for the effective application of CTT is in the range of 300-500. The minimum sample size of 250 examinees proposed to obtain the parameters of IRT. But a sample of 500 examinees is recommended to decrease the measurement error.

The three sub-test of internal medicine, pediatric diseases and General surgery with 45, 25, 24 items respectively was considered. To equal the number of items using simple random sample, 20 items were obtained for each sub test.

Three different sampling plans were used to estimate item difficulty parameter under the CTT and IRT in each of three sub-tests. Two random samples of 500 examinees of each sub-test of internal medicine, pediatric diseases, and general surgery were obtained based on simple random sampling method. Also, both male and female sample of 500 examinees of each of these sub-tests were obtained from the statistical community. Finally, two truncated of high ability and low ability group of 500 examinees were obtained so that the examinees whose total score belonged to 24% of highest scores were classified in high ability group and those whose total score belonged to 24% of lowest scores were classified in low ability group. These sampling plan generated samples that were different in terms of performance on the tests. The high-ability group was defined as those whose total scores fell within the 66% to 100% range and the low ability group was defined as those whose total scores fell within the 0 to 24% range of each three sub test.

The degree of invariance of item difficulty parameter was assessed by correlating item difficulty parameter estimates of two different samples within each measurement framework. The three different sampling plans permitted to assess of item difficulty invariance across increasingly dissimilar samples: (a) between two random samples of the same population, (b)

between female and male samples, and (c) between high and-low-ability samples.

The CTT analysis was conducted by means of SPSS version 17 to estimate the item difficulty parameters under CTT framework, the proportion of correct response to particular item was computed.

The IRT analysis was conducted by means of winestep version 3.71 by joint maximum likelihood (JML) method with Expectation maximation and 24 iterations and 0.05convergence criteria. The Pearson correlation was used to investigate the item difficulty invariance within the same sampling plan, within the same measurement framework (i.e., IRT to IRT, CTT to CTT) across each sub-test. To correct the bias in the sample correlation coefficient, fisher correlation was used to investigate item difficulty invariance within the same sampling plan within the same measurement framework (i.e., IRT to IRT, CTT to CTT) across each sub-test.

The following ethical issues were considered in the research: After obtaining permission from the authorities, the educational program began at the Department of Shiraz University of Medical Sciences. At the start after introduction, the researcher explained the research purposes and the need for its implementation to participants and obtained informed written consent from them. The participants were assured that all information collected will remain confidential. Moreover, considering ethical issues and security, the samples were asked to refrain from writing their name and surname in the questionnaire. They were also assured that the information will only be investigated by the researcher, and only the coding method was utilized for questionnaires to track questionnaires before and after education.

Results

Table 2 and 3 present the results addressing the research question how invariant are the CTT based and IRT-based item difficulty estimates? By analyzing the comparability of average correlation between item difficulty estimates from two different samples derived from same measurement framework in 500 examinees and 20 items for each sub-test.

Table 2 indicates that CTT item parameters were strong invariant for the random sampling plan with correlation ranging from/ 0.986 to, 0.991. For gender sample plan, CTT item parameters showed signs of strong invariance with correlations ranging from, 0.986 to 0.987. The ability sample plan showed opposite results to other sampling plans. There was no invariance between CTT difficulty estimates in performance groups. Nearly all of the correlations generated using the fisher correlation matched with Pearson correlations, however, the largest disagreement between the correlations was 0.114.

Table 2. Invariance of item difficulty parameter from CTT measurement framework: Average between sample correlations of CTT item difficulty parameters (number of examinees = 500) (number of items = 20)

CTT model			
Sampling frame	Sub-tests	P values	Fisher correction P values
Random Samples			
	Internal medicine	0.986(p<0.001)	0.987
	General surgery	0.990(p<0.001)	0.990
	Pediatric diseases	0.991(p<0.001)	0.991
Female-Male Samples			
	Internal medicine	0.961(p<0.001)	0.963
	General surgery	0.961(p<0.001)	0.963
	Pediatric diseases	0.986(p<0.001)	0.987
Samples			
	Internal medicine	-0.098(0.682)	-0.100
	General surgery	0.111(0.640)	0.114
	Pediatric diseases	0.118(0.621)	0.121

Note: The p-values are presented in parentheses. Pearson correlation was used to investigate item difficulty parameters invariance across different samples. Fisher correlation was used for bias.

Table 3 shows that IRT-based item difficulty parameters were also strong invariant, with correlations ranging from. 987 to 990 for the Random sampling plan. For gender sample plan, the IRT-based item difficulty estimates also indicate strong signs of invariance, with correlations ranging from .954 to .987.

Ability sampling plan displays contrary results to other sampling plans. There was no invariance between IRT difficulty parameters in performance groups. Nearly all of the correlations generated using the fisher correlation matched with Pearson correlation. However, the largest disagreement between the correlations was. 0.003.

Table 3. Invariance of item difficulty parameter from IRT measurement framework: Average between sample correlations of IRT item difficulty parameters (number of examinees=500) (number of items=20)

IRT model			
Sampling frame	Sub-tests	1 P	Fisher correction 1 P
Random Samples			
	Internal medicine	0.987(p<0.001)	0.988
	General surgery	0.992(p<0.001)	0.992
	Pediatric diseases	0.990(p<0.001)	0.990
Female-Male Samples			
	Internal medicine	0.957(p<0.001)	0.959

	General surgery	0.954(p<0.001)	0.956
	Pediatric diseases	0.987(p<0.001)	0.988
High-low ability Samples	Internal medicine	-0.170(0.473)	-0.174
	General surgery	-0.078(0.745)	0.035
	Pediatric diseases	-0.175(0.460)	0.179

Note: The p values are presented in parentheses. Pearson correlation was used to investigate item difficulty parameters invariance across different samples. Fisher correlation was used for bias. 1P=one-parameter

Discussion

This qualitative empirical comparison attempted to answer the question of invariance property of item difficulty parameter based on CTT and IRT across different group of examinees. The findings suggest that across samples the degree of invariance of the CTT item difficulty index was very similar with IRT item difficulty estimates. The result of our study supported previous findings by these authors [19-21]

Overall, the findings of this study didn't demonstrate the superiority of invariance property of item difficulty parameter based on IRT to CTT. The similarity of invariance property of item difficulty estimate was quite obvious in two measurement framework.

The results of this study are part of developing body of literature that supports Thorndike's Skepticism (1982) made the following comment with regard to IRT measurement framework: for the large bulk of testing, both with locally developed and with standardized test, I doubt that there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting test will continue to have much the same properties" [22].

Acknowledgements

The authors thank the Vice-chancellor of Shiraz University of Medical Sciences for supporting this research (Code: 5568). This manuscript is extracted from the thesis of Majid shafiayan. The authors also thank the Clinical Research Development Center of Shiraz University of Medical Sciences for the statistical analysis. This paper written based on financial support of Deputy of Research of Shiraz University of Medicals Sciences as a thesis in partial fulfillment of the requirements for the degree of M.Sc. in medical education. We are also pleased to thank the office of graduate studies of Shiraz University of Medical Sciences and Medical Educational Measurement Center of Ministry of Health of IRAN.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was provided for this work.

References

1. Van der linden, A., & Humbleton, R. (1980). Introduction to scaling: New York: Wiley.
2. Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores: MA:Addison-Wesley.
3. Kaplan, R. M., & Saccuzo, D. P. (1997). Pschological testing:Principles,applications and issues: Pacific Grove:Brooks Cole Pub.Company.
4. Adedoyine, O. O., Nenty, H. J., & Chilisa, B. (2008). Invistigating the invariance of item difficulity parameter estimates based on CTT and IRT. Educ Res, 3(2), 83-93
5. Rezaei R, Mehrabani G. A comparison of the scorings of real and standardized patients on physician communication skills. Pak J Med Sci. 2014 May; 30(3):664
6. MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person parameters based on item response theory versus classical test theory. Educ Psychol Meas, 62, 921-943
7. Zarafsanaiey N. Amini M, Saadat F.A comparison of educational strategies for the acquisition of nursing student's performance and critical thinking: simulation-based training vs. integrated training (simulation and critical thinking strategies), BMC Med Educ (2016) 16:294, DOI 10.1186/s12909-016-0812-0.
8. Rezaee R, Shokrpour N. Performance assessment of academic departments: CIPP model. Euro Journal of Social Sciences. 2011; 23(2):227-36.
9. Zarshenas L, Keshavarz T, Momennasab M, Zarifsanaiey N. Interactive Multimedia Training in Osteoporosis Prevention of Female High School Students: An Interventional Study. Acta Med Iran. 2017 Aug 1; 55(8):514.
10. Fan, X. (1998). Item response theory and classical test theory:An empirical comparison of their item/person statistics. Education and Psychological Measurement, 58, 357-381
11. R. K., Swaminathan, H., & Rogers, H. j. (1991). Foundamental of item response theory: Newbury Park,CA:Sage.
12. HambletonLinacre, J. M. (2002). What do Infit and Outfit,Mean-Square and Standardized Mean. from http://www.rasch.org/rmt/rmt162f.htm

13. Progar, S., & Socan, G. (2008). An empirical comparison of item response theory and classical test theory. Horizones of Psychology, 17(3), 5-24
14. Sirotnic, K. A. (1987). The Information Side of Evaluation for Local School Improvement. Int J Educ Res, 11(1), 77-90
15. Courvile, T. G. (2005). Am empirical comparison of item response theory and classical test theory item/person statistics? A&M University.
16. Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. Int J Educ Psychol Assess, 1(1), 1-11
17. .Maier, k., s. (2001). A Rash hierarchial measurement model. J Educ Behav Stat, 26, 307-331
18. Hambleton, R. K., & Swaminathan, H. (1985). Item response theory.Principales and applications. Boston: MA:Kluwer Academic Publishers.
19. Adegoke BA. Comparison of Item Statistics of Physics Achievement Test using Classical Test and Item Response Theory Frameworks.
20. Adedoyin OO, Nenty HJ, Chilisa B. Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. Educational Research and Reviews. 2008 Mar 1; 3(3):83.
21. Courvile, T. G. (2005). Am empirical comparison of item response theory and classical test theory item/person statistics? A&M University.
22. Thorndike, R. l. (1982). Educational measurement:Theory and practice.