International Journal of Information Technology and Business Management
27th May 2012. Vol.1 No. 1

ISSN 2304-0777 www.jitbm.com

INVESTIGATE INTO INVARIANCE PROPERTIES OF ITEM RESPONSE THEORY (IRT) BY TWO AND THREE PARAMETER MODELS

G. Mallikarjuna,
Sikkim Manipal University-DE, India

Dr. V. Natarajan,
Professor Emeritus DEL, MeritTrac, India

Email: **mallikarjuna.g@smudde.edu.in** **drvnatarajan@merittrac.com**

ABSTRACT

Despite the well-known theoretical differences between item response theory (IRT) and classical test theory (CTT), research examining their empirical properties has failed to reveal consistent, demonstrable differences due to usage of samples with errors, incorrect assumptions & incorrect calculations. Dr. Natarajan (India, 1984) had proved the item invariance with his classic 20 X 76 test data and group invariance with 25 X 1000 test data. However, this research will investigate the invariance property of IRT using real live data and also find out how far is this IRT theory an enabler of accurate estimation of parameters over and above the classical test theory (CTT) properties? The objectives of the present research will be: a) to demonstrate through an illustration, the classical test theory (CTT) is a group dependent statistical analysis and bring out its limitations b) to ascertain group invariance of item parameters and item invariance of ability parameters in IRT c) to hypothesize that IRT based item parameter/ability parameters are invariant and more accurate than CTT characteristics. A large scale country wise computer based test (CBT) examinations assessment data is used for this research. The current findings indicate that in CTT, the facility value and index of difficulty are group dependent. Studies are being carried out on invariance property, the core and critical characteristic of IRT for which results are expected in next one year.

KEY WORDS: *Classical test theory, facility value, index of difficulty, item response theory, Item parameters, ability parameters, invariance…*

1. CLASSICAL TEST THEORY

"Classical Test Theory, popularly known as CTT, started off as majority of practices developed during the 1920's.This theory has component theories like Theory of Validity, Theory of Reliability, Theory of Objectivity, Theory of Test Analysis, Theory of Item Analysis etc. (Gulliksen, 1950; Lord & Novick,1968; Dr.Natarajan, 2009) CTT is best suited for traditional testing situations, either in group or individual settings, in which all the members of a target population

are administered the same or parallel sets of test items, for instance, test takers seeking admission in a college or recruitment to a job. These item sets can be presented to the test taker in either a paper-and-pencil or a computer format. Regardless of the format, it is important for the measurement of individual ability that the items in each item set have "difficulties" that match the range of ability or proficiency in the population.

In addition, precise estimation of individual ability requires the administration of a "large enough" number of items whose difficulty levels narrowly match the individual's level of ability or proficiency. For heterogeneous populations, these requirements of the "fixed length" test result in an inefficient and wasteful testing situations that are certainly frustrating to the test taker and not very valid and reliable from the test administrator's and analyst's point of view. Despite its popularity, CTT has a number of shortcomings that limit its usefulness as a foundation for modern testing. The emerging role of computing technology in mental testing highlights some of these limitations of CTT." An advantage with CTT is that it relies on weak assumptions and is relatively easy to interpret.

2. ITEM RESPONSE THEORY:

The approach based upon items rather than test scores was known as Item Response Theory (IRT). While the basic concepts of IRT were, and are, straightforward, the underlying mathematics was somewhat advanced compared to that of CTT. It was difficult to examine some of these concepts without performing a large number of calculations and advancement in computer technology had accelerated the development of IRT. Although difficult to implement in practice, IRT is the formulation of choice for modern testing.

2.1 Contributions in the Area of IRT:

Over the past century, many persons have contributed to the development of IRT. D.H. Lawley (1943) had published a paper on item characteristic curve (ICC) showing that many of the constructs of CTT could be expressed in terms of ICC parameters. Dr. F. M. Lord of the Educational Testing Service had systematically defined, expanded and explored the theory as well as developed computer programs needed to put the theory into practice. His works have been the driving force behind both the development of the theory and its application for the past 50 years. Dr. B.D. Wright (late 1960s) of the University of Chicago recognized the importance of the measurement work done by the Danish mathematician Georg Rasch. He had played a key role in bringing IRT (Rasch model), to the attention of practitioners. Frank Baker (1985) came out with an introductory text book on IRT along with software for the Apple II and IBM personal computers. This program freed the readers from the tedious statistical calculations.

Dr. Natarajan (India, 1984) was the pioneer who introduced IRT to India by publishing a book "Monograph on sample free item analysis" which addressed all three models of Rasch, Birnbaum & Fred Lord. He was awarded a D.Litt by Pune University, India for his thesis An Application of Item Response Theory to Aid Discrimination Function in Achievement Testing. He was instrumental in implementing IRT analysis for finalizing the merit lists in admissions tests conducted by several leaders in India like MIBE, AIIMS, CMC, and REC.

2.2 Basic Concepts Of IRT:

The basic concepts of IRT include Ability, Difficulty and True Score.

2.2.1 Ability: Latent traits such as scholastic, reading, mathematical, and arithmetic abilities are termed as ABILITY which is easily described and its attributes can be listed but cannot be measured directly as the variable is a concept rather than a physical dimension.

In CTT, the number right score on a multiple choice test is used to indicate what the ability of a test taker is. But in IRT, the probability of the correct response to an item is summed up for all items answered correctly in a test indicating the ability of the person taking the test. Each item is dichotomously scored. When the item response is correct, the test taker receives a score of one; an incorrect answer receives a score of zero.

2.2.2 Difficulty: A test based on IRT consists of items that are calibrated for its parameters & different items in a test will have different parameters. Difficulty (item difficulty) is the probability of getting a correct response by a

test taker. For extremely low ability (-∞, -4 or -3) it is 0 or almost 0 & extremely high ability (+∞, +4 or +3) will be almost 1, tending towards 1 but not equal to 1. The ability corresponding to 0.5 probabilities is defined as item difficulty of the item. Thus, the item difficulty of an item and the ability of the test taker are on the same scale and provide a relationship between test items and true scores of test takers.

2.2.3 True Score: In CTT, the true score is the mean of several number right scores of the test taker over the same or equivalent test which is impractical or impossible to obtain.

Standard Error of Measurement (SEM) specifies the limits of any number right score. For example, if a test taker's number right score is 72 and SEM of the test is 7, the test taker's score can range between 65 (72-7) to 79 (72+7) for 2/3rd probability. IRT enables the estimation of True Score (TS) from a test taker's ability, with a low percent of error (usually of the order of 0.1 percent) .The true score for an estimated ability of a test taker (θ) is the sum total of probability of a correct response of all the items (with different item difficulty values), that is, **TS(θ) = Σ of individual probability of correct answers to all items (D.H Lawley)**

2.3 Group Invariance of Item Parameters:

Group invariance of the item parameters says that the values of the item parameters are a property of the item, not of the group that responded to the item. These item parameters can be estimated from any group of test takers & are group invariant. The term group invariance refers to this independence of the item parameter estimates from the distribution of ability.

2.4 Item Invariance of a Test Taker's Ability Estimate

If all the items measure the same underlying latent trait and values of all the item parameters are in a common metric then the test taker's ability is invariant with respect to the items used to determine it. On the average, estimated ability will be same even if the test taker takes an easy or hard test.

Under IRT, the test taker's ability is fixed and invariant with respect to the items used to measure it.

- If he takes the same test several times assuming he does not remember the items or the responses from test to test then his ability would be fixed.
- However, if he received remedial instruction between the tests or if there were carryover effects, his underlying ability level would be different for each testing.

- Thus, the test taker's underlying ability level is not immutable.

3. ITEM RESPONSE THEORY OVER CLASSICAL TEST THEORY

IRT provides relationships between item parameters and the ability of the test taker .It provides adaptable and effective methods of test construction, analysis and scoring than those derived from CTT. For a CTT test to attain the comparable accuracy of IRT, it requires 200 items and more. In IRT it needs small number of items and hence a small item bank.

IRT scale scores are functions of estimated item parameters - without altering the interpretation of the test scale, items can be retired and replaced .The scoring in IRT absorbs possible differences in the characteristics (difficulty, discriminating power etc.) between the retired items and the replacements and there is no need to find new items with the same difficulty and discriminating power as the old items or for an equating study of the revised test separate from its operational use, as required in CTT.

Unique property of IRT is the location of items and the test takers on the same scale. It enables to state the probability that a test taker at a particular score level will answer a given item correctly. IRT permits the "content referencing" of the scale scores. Under CTT, the test taker's raw test score would be the sum of the scores received on the items in the test.

The basic concepts of IRT rest upon the individual items of a test rather than upon some aggregate of the item responses such as a test score. The main aim is whether a test taker got each individual item correct or not, rather than in the raw test scores.

All the above content taken with permission from the electronic book of Dr. V. Natarajan entitled Basic principles of IRT and application to practical testing and assessment, self-published in electronic media, 2009.

4. THE PURPOSE OF PRESENT RESEARCH INVESTIGATION

Most of the research done over the years on invariance property of IRT, the core and critical characteristic is based on theoretical analysis and statistical calculations on simulation data. A major part concerning the theoretical work was produced in the 1960's (Rasch, Birnbaum; Lord & Novick). Empirical studies examining the invariance characteristics of item and person statistics of IRT are very scarce. Very few investigations have been done with the usage of real data and such investigations except few could not establish the invariance property due to usage of samples with errors, incorrect assumptions & incorrect calculations.

Dr Natarajan (India, 1984) had proved the item invariance with his classic 20 X 76 test data and group invariance with 25 X 1000 test data. O. O. Adedoyin, H. J. Nenty and B Chilisa (2008) and O.O. Adedoyin(2010) studies supported the invariance principle of IRT. However, studies by Fan (1998), Lawson (1991), MacDonald and Paunonen (2002), Skaggs and Lissitz (1986,1988) and Stage (1998a, 1998b, 1999) have all pointed to little difference between item response estimates and classical test theory estimates

Given the limited number of empirical studies directly or indirectly addressing the invariance issue, there is an obvious lack of systematic investigation about the invariance of the item and person statistics obtained from IRT frameworks.
This research has been derived from the need to investigate the invariance property of IRT with the usage of real live data, appropriate sampling techniques, new and recent trends in statistical advancements/calculations along with the latest software's to establish invariance property of IRT and also find out how far is this IRT theory an enabler of accurate estimation of parameters over and above the classical test theory (CTT) properties? Since the population of the subjects to be used in this study is large, it is envisaged that the findings of this research study will be reliable, objective and valid.

The present research is focused on the following: to demonstrate through an illustration, the classical test theory (CTT) is a group dependent

– statistical analysis & bring out the limitations of CTT

– to ascertain group invariance of item parameters and item invariance of ability parameters through statistical calculations, using BILOG MG3 for two parameter & three parameter models on an online examination data of Sikkim Manipal University (February 2012)

– to hypothesize that IRT based item parameters are invariant and more accurate than CTT characteristics

– to hypothesize that IRT based ability parameters are invariant and more accurate than CTT characteristics

5. METHODOLOGY

5.1 Examinee Sample & Data

The data used consists of 6544 students of MBA Distance Education, Sikkim Manipal University, India who had appeared for the online examinations in the subject area of Research Methodology during February 2012. Sikkim Manipal University is one of the largest distance education service providers with all India presence & foot print across 22 countries. The test is computer based consists of 40 multiple choice items which were dichotomously scored.

The following type of sample groups were drawn from the data

i) Top Half- Bottom Half ,each consisting of 3227 examinees

ii) Examinees from two different states, group 1 consisting of 1173 examinees of Uttar Pradesh and group 2 consisting of 1043 examinees from Delhi.

iii) Higher ability –Lower ability (27%) groups consisting of 1767 students.

Hypothesis 1: The item parameters estimated using IRT two as well as three parameter models are invariant and are superior to the statistically

calculated index of difficulty, discrimination index relevant for both the models of IRT.

Hypothesis 2: To establish that ability parameter of each of the test takers is an invariant no matter which group of items like odd numbered items in a test or even numbered items in a test are used and that the ability parameters are more accurate than the corresponding CTT number right scores.

5.2 Facility Value (P) & Index of Difficulty (Q) Under CTT are Group Dependent

Three different types of sample consisting of 2 different examinee data as groups are drawn from the data. The facility value (P=No of persons getting an item right / no attempting it) & Index of difficulty (Q=1-P) were calculated for each sample and compared with other sample of the group.

The percentage difference in the facility value and index of difficulty values is given in Table 1.

Table 1 – Percentage Difference in the Facility Value and Index of Difficulty

Table 1	Top Half-Bottom Half		State 1(UP)-State 2(Delhi)		HAG-LAG	
	P	Q	P	Q	P	Q
Item 1	-20.7	21.6	-16.7	15.1	63.0	-297.4
Item 2	-6.5	13.6	-15.6	23.8	50.8	-563.3
Item 3	-10.1	23.5	-9.7	19.1	46.3	-796.3
Item 4	-8.4	4.3	-4.4	2.1	57.8	-66.5
Item 5	-15.2	6.4	-23.9	8.6	71.2	-84.4
Item 6	-13.0	5.9	-28.9	11.0	73.6	-105.8
Item 7	-22.1	8.5	-18.8	6.6	70.0	-81.9
Item 8	-10.0	15.5	-14.3	18.1	58.6	-488.9
Item 9	-10.8	8.1	-17.7	12.1	60.7	-135.0
Item 10	-13.5	17.4	-28.2	24.9	65.5	-496.5
Item 11	-14.8	8.0	3.0	-1.9	41.0	-45.1
Item 12	-11.4	2.9	-20.4	4.4	54.1	-25.7
Item 13	-11.8	7.1	-24.3	13.2	61.8	-101.1
Item 14	-14.8	12.1	-12.9	8.4	63.8	-183.6
Item 15	-12.8	14.7	-26.7	23.4	56.8	-247.1
Item 16	-10.6	25.5	-29.0	42.4	55.6	-1389.1
Item 17	-11.6	17.1	-30.6	30.9	62.5	-487.1
Item 18	-15.4	6.3	-30.1	10.1	72.5	-76.6
Item 19	0.4	-0.1	0.2	-0.1	-10.8	4.0
Item 20	-15.1	14.2	-23.9	18.8	63.9	-228.8
Item 21	-8.7	24.7	-16.5	35.9	44.9	-638.7
Item 22	-17.4	7.2	-16.0	5.5	59.0	-55.5
Item 23	-28.1	11.3	-12.4	5.2	69.8	-83.0
Item 24	-4.3	1.6	6.0	-2.6	17.3	-8.3
Item 25	-35.7	13.7	-34.1	11.6	83.1	-170.5
Item 26	-28.3	13.0	-35.8	14.4	78.7	-137.8
Item 27	-12.4	21.0	-28.1	31.7	56.1	-476.7
Item 28	-12.0	7.0	-20.1	10.5	60.7	-80.0
Item 29	-15.2	7.2	-13.4	5.8	65.1	-78.0
Item 30	-10.4	10.4	-27.6	23.8	48.3	-126.0
Item 31	-7.5	9.9	-18.0	17.9	52.8	-159.2
Item 32	-6.6	5.5	-9.2	6.5	64.9	-152.7
Item 33	-7.4	16.0	-11.9	19.5	39.8	-274.0
Item 34	-14.5	8.1	-12.7	8.0	43.4	-53.4
Item 35	-11.3	21.5	-32.4	40.1	58.2	-868.2
Item 36	-9.3	5.8	-15.2	8.5	59.5	-97.6
Item 37	-11.3	22.3	-24.1	35.2	56.5	-1078.3
Item 38	-27.6	6.1	-18.1	3.7	67.1	-39.4
Item 39	-10.1	7.8	-8.0	6.6	38.6	-52.9
Item 40	-16.1	14.4	-23.3	15.0	71.3	-332.7

The z-test of significance was carried out for index of difficulty values of all the three types of sample groups at 95% level of significance and the same is given in Table 2.

Table 2 – Significance Values

Table 2	Top Half-Bottom Half	State 1(UP)-State 2(Delhi)	HAG-LAG
Z value	4.54	3.6	23.47

5.3 Person Parameter in CTT is Test Dependent

Two different subset test forms one consisting of 20 odd items and other comprising 20 even items were generated from the test. The person ability was estimated as the sum of each examinees response to all the test items in each subtest. The number right score (raw score) for all 6544 examinees was found out for each subset of test.

6. RESULTS AND DISCUSSION

To find out whether or not the facility value and index of difficulty estimates based on CTT are invariant across different groups, three types of sample groups consisting of 2 subsets were used. The differences in "P" and "Q" values between samples of each group were transformed to percentiles. The same are shown in Table 1 and the difference is very significant for all groups which can be deduced from the z values obtained at 95% confidence level. Table 3 shows the percentage difference of minimum and maximum P & Q values for each group.

Table 3 – Percentage Difference of Minimum and Maximum P & Q Values of each Group

Table 3	Bottom Half-Top Half		State 1(UP)State 2(Delhi)		HAG-LAG	
	Percentage Difference					
	P	Q	P	Q	P	Q
Min	0.4	0.1	0.2	0.1	10.8	4
Max	35.7	25.5	35.8	42.4	83.1	1389.1

To find out whether or not the person parameter estimates based on CTT are test dependent, 2 test subsets comprising of 20 Odd and 20 even items were used for estimating the person parameter. The raw score (number right score) of 4138 examinees was found to be different in the test subsets. Hence, subsets of test items developed to measure the same ability have significant influence on the estimate of such ability for the examinees. This implies that the examinees score or ability is dependent on the particular set of items administered (i.e.) it is test-dependent. The person parameter estimates were found to vary cross the item groups from the same test.

The findings of the present study **(Part 1 of the research)** are, in CTT the facility value & index of difficulty are group dependent and the person parameter is test dependent.

In the next part of research IRT properties of group invariance of item parameters and item invariance of ability parameters are to be demonstrated with the sample taken from the above CTT analysis. For that, the methodology that this research has proposed to adopt is

- to perform a comprehensive test and item analysis with the help of BILOG MG3 by way of outputs giving item parameter values, ability values and true scores of test takers. Establish the accuracy of IRT and show non dependence on group for its values using 2 and 3 parameter models.

- with the help of responses from SMU online exams, establish that there is a group invariance of item parameters using two groups of different states student data , HAG- LAG (27%) and top half - bottom half. Similarly, with the help of the same prove the item invariance of the ability parameter.

7. REFERENCES:

1. Adedoyin. (2010). Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories. *International Journal of Educational Science*, 107-113.

2. Adedoyin, J.Nenta, H., & Chilisa, B. (2008). Investigating the Invariance of Item Difficulty Parameter Estimates Based on CTT and IRT. *Educational Research and Review Vol. 3*, 083-093.

3. Baker, F. B. (2001). *The Basics of Item Response Theory.* USA: ERIC Clearinghouse on Assessment and Evaluation.

4. Cikrikci-Demirtasli, N. (2002). A study of Raven Standard Progressive Matrices Test's Item Measures Under Classic and Item Response Models: An Empirical Comparison. *Journal of Faculty Educational Sciences*, 1-2.

5. Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics.* Texas: Graduate Studies of Texas A&M University.

6. Dr.V.Natarajan. (1984). *Monograph on Sample Free Item Analysis.* published by AIU.

7. Dr.V.Natarajan. (2009). Basic Principles of IRT and Application to Practical Testing.

8. Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement*, p357(25).

9. Galdin, M., & Laurencelle, L. (2010). Assessing parameter invariance in item response theory's logistic two item parameter model: A Monte Carlo investigation. *Tutorials in Quantitative Methods for Psychology*, 39-51.

10. Kelkar, V., Wightman, L. F., & Luecht, R. M. (2000). Evaluation of the IRT Parameter Invariance Property for the MCAT. *Annual Meeting of the National Council on Measurement in Education* (pp. 25-27). Greensboro: University of North Carolina.

11. Lourdes, M., & Franco, M. (n.d.). The philippine aptitude classification test: why shift from classical test theory to item response theory. *center for Educational Measurement*, 1-16.

12. Magno, C. (2009). Demonstrating the Difference be-tween Classical Test Theory and . *The International Journal of Educational and Psychological Assessment*, 1-11.

13. Progar, S., & Socan, G. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. *Educational and Psychological Measurement*, 5-24.