Comparing performance of statistical models for individual's ability index and ranking

Iqbal, Javed
International Islamic University (IIU), Islamabad, Pakistan

01. January 2012

Comparing Performance of Statistical Models for Individual's Ability Index and Ranking

Javed Iqbal
Ph.D Scholars (Econometrics)

International Islamic University Islamabad

Abstract

Efficient allocation of resources is the basic problem in economics. Firms, educational institutions, universities are faces problem of estimating true abilities and ranking of individuals to be selected for job, admissions and scholarship awards etc. This study will provide a guide line what technique should to be used for estimating true ability indices and ranking that reveals ability with maximum efficiency as well as it clearly has the advantage of differentiating among individuals having equal raw score. Two major theories Classical Testing Theory and Item Response Theory have been using in the literature. We design two different Monte Carlo studies to investigate which theory is better and which model perform more efficiently. By discussing the weaknesses of CTT this study proved that IRT is superior to CTT. Different IRT models have been used in literature; we measured the performance of these models and found that Logistic P2 model is best model. By using this best model we estimate the ability indices of the students on the basis of their entry test scores and then compare with their abilities obtained from final board examination result (used as proxy of true abilities). This is a reasonable because the final exam consists of various papers and chance variation in ability Index is a minimum. With real life application this study also proved that IRT estimate the true abilities more efficiently as compared to classical methodology.

Key words: Ability Index, Monte Carlo study, Logistic and Probit models, Item Response Theory, Classical Test Theory, Ranking of Students

1. Introduction and Motivation:

Ranking of individuals according to their abilities is a problem of great interest for people

---

Correspondence Address: Javed Iqbal, International Islamic University, Islamabad. Email: javed.chughtai@gmail.com

belonging to all spans of life. For examples, admission of students in an educational institute, selection of candidate for job in different fields and scholarship award for PhD programs etc based upon the ability ranking in the competent exam. We consider the case where ability of individual is reflected in form of discrete response. An example of such response is score of examinee in multiple choice tests. Classical procedure of ranking for an individual, every item is considered as an independent draw for some certain distribution which is a function of ability of individual. However, the assumption of independent draw is questionable. There are several problems in Classical procedure of ranking. For example, in Standardized test for the selection for job, examinees are ranked according to their total scores. However, in the test there are some easy questions and some difficult but examinees attempting a difficult question and an easy question gets equal credit.

Second problem is that in some tests (i.e. GRE, GMAT etc) the questionnaires are divided into multiple categories e.g. analytical, quantitative and verbal etc but examinees belong to different fields such as biology, chemistry, mathematics etc. In some cases more questions are favorable for mathematicians as compare to biologist but all questions will be given equal weights for each examinee.

Third problem with classical procedure is that some times in a test two or more candidates get equal marks creating a problem for deciding their ranks (Zaman & Atiq, 2008(unpublished)).

Fourth problem is that, Classical test statistics are sample dependent in that as the sample changes, the estimators would change (Cantrell, 1997; Henson, 1999). Therefore, the classical test theory estimators are not generalizable across populations (Courville, 2004)

With all these above problems we cannot estimate the true ability indices of the candidates efficiently for the selection for different fields such as scholarship award, job selection, and admission of students in different educational fields. We can select less capable candidate which leads us to allocate resources to less efficient way.

Item response theory (IRT) is an attempt to adjust for problems highlighted above. IRT methods have a long history dating back to Lord (1952) and earlier, and are now well established in the field of education, medical and psychometrics etc, providing an alternative to the classical

testing theory (CTT). IRT is a model-based approach that provides additional tools for measuring traits and abilities by clearly separating test items, characterized by individual item parameters (difficulty parameters, discrimination parameters and guessing parameters etc) from the characteristics of examinees. IRT gives different weights to each item (question) according to their difficulties and discriminations to measure true ability of the candidates.

The key to IRT is a model that links the characteristics of a given item, and the latent ability of an individual subject, to the probability that the subject will respond correctly to that test item. IRT used various types' econometric models such as Logistic models and Probit models for both binary and polytomous test items. The most commonly adopted method for estimating IRT parameters is marginal maximum likelihood (MML), based on the work of Bock & Lieberman (1970) and Bock & Aitkin (1981). Other estimation methods have been reviewed by Baker (1992) and Terry & Ackerman (1993).

Existing literature focuses on technical and theoretical comparison of IRT with classical models that is relevant for professional purposes e.g. an educationist is interested in knowing the quality of test he designed. However, focusing on the problem of ranking of individuals, there is not available any quantifiable measure of relative efficiency of IRT onto CTT. Moreover in IRT there is more than one statistical model based on certain assumptions. For real data set we know nothing about validity of assumptions. Therefore it is needed to investigate a model which is more robust to failure of these assumptions.

2. Objectives:

Ranking of individuals according to their abilities is a problem of great interest for people belonging to all spans of life. There is huge literature on the estimation of true abilities of the candidates appearing in the exam. Many econometrics and statistical models are applied to estimates person abilities parameters and item parameters. Existing literature focuses on technical and theoretical comparison of IRT with classical models. However, focusing on the problem of ranking of individuals, no quantifiable measure of relative efficiency of IRT and CTT is available. There is a little literature on the comparison of models used for estimating the abilities (Courville 2004, Sheng 2005). However, the literature focuses on technical detail which is relevant for professionals, but comparison of the efficiency of the models is not available in the literature. Hence this study is focusing on the following objectives.

- *To measure relative efficiency of classical and IRT models where true ability is assumed to be known.*

People are using IRT in practice as well but how much IRT can improve over Classical model; No quantifiable measure is available to date. We propose Monte Carlo experiment to investigate that IRT can approximate true ability ranking of students efficiently than CTT, also we will provide measure of efficiency.

- *To find out IRT model which is more robust to model's assumptions*

There are certain types of IRT models based on certain assumptions e.g. 2-p models assume that probability of correct response for a certain individual depends on item difficulty and discrimination whereas 1-p model is based on the assumption that probability of correct response is monotonic function of item difficulty only. For the real data sets, we know nothing about validity of assumptions and hence we need a procedure which is more robust to failure of model assumptions.

- *By using real data sets (entry test result of candidates) we will estimate the true abilities and ranks of the individuals by using classical test theory and item response theory and will compare with their final exam's results.*

3. Literature Review:

The relevant literature can be divided into two types; (i) the literature on classical methods (classical testing theory, CTT) for measuring the ability of the candidate appear in the competitive exam. (ii) The literature on item response theory (IRT).

3.1. Classical Testing Theory:

The proficiency (or ability) of a person is often estimated using the number of correct score to the items in the test (or simply the test score). A test score that is equal to a cut-off score or greater than a cut-off score is considered a pass; otherwise a failure. This approach of using the test score as proficiency estimate is sometimes referred to as the classical test theory (CTT) approach (Sotaridona, et. al. 2003)

Classical test theory has served test development well over several decades. Most psychologists should, and in fact do, know its principles however, since Lord & Novick's (1968) classic book introduced, model-based measurement, a quiet revolution has occurred in test theory

(Wiberg, 2004). The major advantage of CTT is its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Russel, 1993). Relatively weak theoretical assumptions not only characterize CTT but also its extensions (e.g., generalizability theory). Although CTT major focus is on test-level information, item statistics (i.e., item difficulty and item discrimination) are also an important part of the CTT model (Fan, 1998). 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. The major limitation 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). Despite the theoretical weakness of CTT in terms of its circular dependency of item and person statistics, measurement experts have worked out practical solutions within the framework of CTT for some otherwise difficult measurement problems. A major challenge for classical test theory is how to score individuals who complete different versions of a test (Fan, 1998). By focusing on the item as the unit of analysis, IRT effectively solved this problem.

3.2. Item Response Theory:

Item response theory (IRT) is an attempt to adjust for problems facing by the classical methods. IRT methods have a long history dating back to Lawley (1943) and Lord (1952). They established the basic concept of item response theory. A major part concerning the theoretical work was produced in the 1960's (Wiberg, 2004).

One of the basic assumptions in IRT is that the latent ability of individuals is independent of the content of a test. The relationship between the probability of answering an item correctly and the ability of individuals can be modeled in different ways depending on the nature of the test (Hambleton et al., 1991). It is common to assume unidimensionality, i.e. that the items in a test measure one single latent ability. According to IRT, test-taker with high ability should have a high probability of answering an item correctly (Wiberg, 2004). Another assumption is that it does not matter which items are used in order to estimate the test-takers' ability. This assumption makes it possible to compare test-takers' result despite the fact that they

have taken different versions of a test (Hambleton & Swaminathan, 1985). IRT has been the preferred method in standardized testing since the development of computer programs. The computer programs can now perform the complicated calculations that IRT requires (Linden & Glas, 2000).

The most commonly adopted method for estimating IRT parameters is marginal maximum likelihood (MML), based on the work of Bock &Lieberman (1970) and Bock and Aitkin (1981). Other estimation methods have been reviewed by Baker (1992), and an MCMC approach has been described by Patz & Junker (1999). Thomas & Cyr(2002) used the three parameters Logistic model and discussed various IRT issues, including point and variance estimates of item parameters, the potential for bias due to ignoring survey weights, biases in the distribution of ability predictors, and the dependence of this bias on test length. Another issue highlighted by Atiqe & Zaman (2008) that in CTT two students with equal raw score have the same ranking while in IRT they have different ranking, making the job of policy maker easier to take decision. Item Response Theory can reduce the problems faced by CTT, therefore the comparison between these methods is necessary.

3.3. Comparison between CTT and IRT (ML estimators):

Over the past twenty-eight years, since Lord's 1980's *Testing Problems*, item response theory (IRT) has become the jewel of large-scale test construction programs. However, some investigations (Fan, 1998; MacDonald & Paunonen, 2002) have studied the empirical difference between these two models. Fan (1998) noted that "Because IRT differs considerably from CTT in theory, and commands some crucial theoretical advantages over CTT, it is reasonable to expect that there would be appreciable differences between the IRT and CTT-based item person statistics" (p. 360).

However, Fan (1998), MacDonald & Paunonen (2002), Stage (1999) etc have indicated a little difference between item response and classical test theory estimates. In Stage's (1999) work with the SweSAT test READ, she noted that the agreement between results from item-analyses performed within the two different frameworks IRT and CTT was very good. It is difficult to find greater invariance or any other obvious advantages in the IRT based item indices. Courville (2004) showed in his comparison study that "CTT and IRT may produce very similar results in a

single test administration. But because CTT estimates are theoretically sampled dependent, across different samples item response theory should yield results that are more invariant."

Another latest comparison study by Jimelo & Silvestre (2009) showed that both measurement methodologies produced very similar item and person statistics both in terms of the comparability of item and person statistics, difficulty level of items, internal consistency and differential item functioning between the two frameworks. With his finding, author asks a very interesting question that how to view the differences between IRT and CTT models both in theory and in testing practice.

Many studies have been conducted to investigate the comparability of items and person statistic. Theoretically CTT is simple and easy to apply, therefore classical test statistics are still commonly used in test construction process, however many researchers have questioned their utility in the modern era. Existing literature focuses on technical and theoretical comparison of IRT with classical models and many studies claimed that IRT works well as compare to CTT but how much IRT can improve over Classical model; No quantifiable measure is available to date. To estimate the true ability of an individual there have been used various IRT models based on certain assumptions. But for the real data sets, we know nothing about validity of assumptions and hence we need a procedure which is more robust to failure of model assumptions. As there is no quantifiable comparison between IRT models in the existing literature we will provide Monte Carlo study to measure the efficiencies of IRT models.

4: Methodology

This study is focusing on three different issues. First we will compare CTT models with IRT models. In second issues, since different models have been using to estimate true abilities under both theories (CTT and IRT) will compare the performance of these models. For these both objectives we will design a Monte Carlo study. In third issues we will measure abilities of the individuals by using best model and the result of test consists of multiple choice questions and compares their abilities with the final board exam result which is used as proxy of true ability of the students. This is reasonable because of the final exam consists of various papers and chance of variation is minimum in this exam. Before discussing the methodology for the design of Monte Carlo study, we are presenting the introduction of these models.

4.1. CTT and IRT models:

To estimate the abilities of the candidates in a competent exam various econometrics models has been used. Following is the brief introduction of these models.

4.1.1. Mo: Classical testing theory (CTT) model.

This is very simple model used to measure the ability θ_i of *ith* candidate in an exam by accounting the total marks in the exam, the student get more marks is considered to be more able as compared to the student who gets less marks.

4.1.2 Item response theory Models.

The unidimensional IRT model provides a fundamental framework in modeling the person-item interaction by assuming one latent trait. Suppose a test consists of k dichotomous (0-1) items, each measuring a single unified trait, θ. Let $Y = \left\lfloor y_{ij} \right\rfloor_{n\times k}$ represent a matrix of n examinees' responses to the k items, so that y_{ij} is defined as $y_{ij} = 1$ if person i answer item j correctly and $y_{ij} = 0$ if person i answer item j incorrectly. For i=1, 2 ... n and j= 1, 2 …k

In Item response theory, person ability parameter θ_i depends upon the item's parameters e.g. α_j (item difficulty parameter), β_j (discrimination parameter) etc. Since we are focusing this study on discrete binary response, we will use probability models instead of linear regression models. Following are the models which have been used in Item response theory to estimate the maximum likelihood estimators of true ability.

i) **M1: Logistic one parameter model.**

The probability of person i obtaining correct response for item j can be defined as

$$P(\theta_i) = P(X_{ij} = 1/\theta_i, \alpha_j) = \frac{\exp(\theta_i - \alpha_j)}{1 + \exp(\theta_i - \alpha_j)}$$

Where α_j is associated with item difficulty, θ_i is a scalar latent trait parameter (ability parameter) and $< -\infty < \theta_i < \infty$ and $-\infty < \theta_i < \infty$ (Baker, 1992; Robert et al., 1988)

ii) **M2: Logistic two parameters model.**

$$P(\theta_i) = P(X_{ij} = 1/\theta_i, \alpha_j, \beta_j) = \frac{\exp(\beta_j(\theta_i - \alpha_j))}{1 + \exp(\beta_j(\theta_i - \alpha_j))}$$

Where α_j is associated with item difficulty, β_j is a slope parameter (discrimination index), θ_i is a scalar latent trait parameter (ability parameter) and β_j >0 (Bock and Aitkin, 1981; Robert et al., 1988)

iii) **M3: Logistic three parameters model.**

$$P(\theta_i) = P(X_{ij} = 1/\theta_i, \alpha_j, \beta_j, \gamma_j) = \gamma_j + (1 - \gamma_j)\frac{\exp(\beta_j(\theta_i - \alpha_j))}{1 + \exp(\beta_j(\theta_i - \alpha_j))}$$

Where α_j is associated with item difficulty, β_j is a slope parameter (discrimination index), γ_j usually referred to as the guessing parameter and θ_i is a scalar latent trait parameter (ability parameter), (Sotaridona et. al, 2003)

iv) **M4: Probit One parameter model.**

For the one-parameter normal ogive (1PNO) model, the probability of person i obtaining correct response for item j can be defined as

$$P(\theta_i) = P(X_{ij} = 1/\theta_i, \alpha_j) = \Phi(\theta_i - \alpha_j) = \int_{-\infty}^{\theta_i - \alpha_j} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt$$

Where α_j is associated with item difficulty, θ_i is a scalar latent trait parameter, and the term\one-parameter" indicates that there is one item parameter α_j in the model. (Sheng, 2005)

v) M5: Probit two parameters model.

For the two-parameter normal ogive (2PNO) model, the probability of person i obtaining correct response for item j can be defined as

$$P(\theta_i) = P(X_{ij} = 1/\theta_i, \alpha_j, \beta_j) = \Phi(\beta_j\theta_i - \alpha_j) = \int_{-\infty}^{\beta_j\theta_i - \alpha_j} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt$$

Where β_j is a positive scalar parameter describing the item discrimination (Sheng, 2005).

vi) M6: Probit three parameters model.

For the three-parameter normal ogive (3PNO) model,

$$P(\theta_i) = P(X_{ij} = 1/\theta_i, \alpha_j, \beta_j, \gamma_j) = \gamma_j + (1-\gamma_j)\Phi(\beta_j\theta_i - \alpha_j) \quad \text{Where } 0 \le \gamma_j < 1$$

Where γ_j j is a pseudo-chance-level parameter, indicating that the probability of correct response is greater than zero even for those with very low trait level (Sheng, 2008).

4.2. Monte Carlo Study to measure the efficiency of CTT and IRT and compare them.

We design a Monte Carlo experiment to measure efficiency of CTT and IRT models and compare them and will see that which theory is better? Our experiment is design as

- We choose a fix vector of abilities θ_i, i = 1, 2… n. where θ_i is taken as true ability of individual i. Since θ_i is normally distributed and theoretical range of ability is from negative infinity to positive infinity however in IRT maximum values of θ_i lies between -3 to 3 (Baker 2001, book) hence we take θ_i ~N(0,1).
- Since theoretical range of the values of difficulty parameter is $-\infty \le \alpha_i \le +\infty$, and typical values have the range $is\ -3 \le \alpha_i \le +3$ (Baker 2001, book).However we takes the value of this parameter as α_j =5*uniform number between (0, 1)-2.5. And value of discrimination parameter is as disc=abs (N (0, 1)) (Atiq and Zaman, 2008). And finally the guess parameter γ_j is uniformly distributed between 0 and 1 as according to Baker 2001. By choosing any one of the Models given above (say M1) for data generating process (DGP) we generate data of responses. In this study we fix the number of individuals' n=100 and number of items are 30.
- Re estimate abilities under CTT and IRT by using same model M1.

- Standardize these IRT and CTT abilities and Compute Mean Absolute Error (MAE) of these abilities with true abilities.
- Repeat this process different time by changing sample of true abilities and takes the averages of MAE of these repetitions and then compare the MAE of abilities measured by CTT and IRT and the model having less MAE is better.
- Repeat this process by choosing second model (say M2) as DGP and so on.
- Compare the MAE of all these models with each other and select the most power full model which has low MAE.

In this experiment we get two results. 1st is the comparison between CTT and IRT and the 2nd is comparison between different models (CTT and IRT models)

4.3. Monte Carlo experiment for the comparisons of different models on the bases of their efficiencies.

In this Monte Carlo experiment we compare all the models used to estimate the true abilities of the individuals. Our experiment is design as,

- We will choose a fix vector of abilities θ_i , i = 1, 2… n.
- Choose any one of the Models given above (say M1) for data generating process (DGP).
- Re estimate the abilities by using all models M1:M6.
- Compute Mean Absolute Error (MAE).
- Repeat this process by choosing second model (say M2) as DGP and so on.
- Compare the MAE of all these models with each other and select the most power full model which has low MAE.

4.4. Real life application.

Monte Carlo study helps us in finding more efficient model to estimate true ability of the test-takers. With the help of this model we estimate the ability of the candidates who take part in the examination (MCQ type entry test). We compare their abilities estimated by using the data of the MCQ test with the final board examination result and will see that is there any improvement by using IRT models as compared to CTT models?

4.5. Tools

For this Monte Carlo study and real life examples we used Matlab software for programming*. For this purpose we developed the programming ourselves and no specialized software for IRT was used. The program coding can be seen in appendix A and B.

5. Empirical Findings

We will discuss the results of these three studies one by one.

5.1. Comparisons between CTT and IRT.

Due to tedious nature of IRT analysis without specialized software we takes a vector of 100 abilities and number of items are 30 in this Monte Carlo experiment. We generate a data set of responses (in the farm of 0 and 1. Where 0 for false answer and for correct answer we take1) by using these models one by one and then re-estimate the true abilities under both CTT and IRT techniques. Mean absolute error of ability indices is given in table 1(A) given below. From the table it is clear that MAE for IRT is less as compared to CTT in all the six models and the smallest mean absolute error is for Probit P2 IRT model.

Table: 1 (A) Comparison between CTT and IRT on the basis of Mean Absolute Error.

Model\Mean absolute error	True Ability Vs CTT	True Ability Vs IRT
Logistic P1	0.40683	0.406712
Logistic P2	0.471838	0.396247
Logistic P3	1.031671	0.884584
Probit P1	0.626708	0.625926
Probit P2	0.376822	**0.285877**
Probit P3	1.051641	0.869906

The main issue in CTT is that two are more individuals with equal raw scores having same ranks while in IRT there is very little chance for two or more candidates having same score in an examination. In table1 (B), we presents the correlation between the true ranks and the ranks of the individuals on the basis of their scores estimated by different methods/models. From table's results we can see again in all the cases, the ranks calculated by IRT are more correlated with true ranks as compared to CTT ranks. Hence IRT estimation of ranking is more reliable than CTT. In this Monte Carlo study, we generate the data with one model by taking true abilities as

* MATLAB codes can be made available on request.

given and re-estimate these abilities with the same model. From the results we can say that Probit P2 model is best as its IRT correlation coefficient is highest (r=.93) as compared to other models however there is not big difference between the CTT correlation and IRT correlation estimated with Probit P2 model. But if we see the table1(B) with another angle we can conclude that Probit P3 model is more efficient because if data is generated and estimated with Probit P3 models then there is big difference between the CTT correlation and IRT correlation. Since for the real data sets, we know nothing about validity of model's assumptions so we cannot say which model is best in this experiment. For this purpose, comparison of models, we will give another Monte Carlo experiment in the next section.

Table: 1(B) Comparison between CTT Rank and IRT Rank on the basis of their correlation.

Model\Correlation	True Ability and CTT	True Ability and IRT
Logistic P1	0.874	0.871
Logistic P2	0.823	0.879
Logistic P3	0.222	0.339
Probit P1	0.698	0.705
Probit P2	0.893	**0.936**
Probit P3	0.054	0.399

Graph1 in Appendix A is the graph of ranks of the individuals according to their abilities estimated by Probit P3 models. We can see that the candidate with CTT rank 1, its true rank is 82 and IRT rank is 88. Student at CTT rank 2 having true rank 71 while IRT estimated rank is 91. Hence Graphical analysis shows that IRT ranks are more correlated as compared to CTT ranks.

5.2. Comparisons between Different IRT models.

From the results of last section it is concluded that Items Response Theory is more reliable than Classical methodology for estimating the true abilities of individuals. Since there are different IRT models with different assumptions have been used in IRT estimation techniques. In this Monte Carlo study we compare the performance of these models on the basis of their efficiencies. We generate the data of responses by taking a vector of true abilities by using one model. After generating the data, we re-estimate the abilities with the help of all other models. Results of the mean absolute error of these estimated abilities are given in the table 2(A).From the table we can see that MAE of Logistic P2 model is minimum in five cases (GDP from five models) out of six as compared to all other models for re-estimation. Only when data is generated by Probit P2 model, Logistic P2 model is at 2nd place of superiority.

Table: 2(A) Comparison between different models for estimation of true ability Indices under IRT on the basis of mean absolute error.

Models for dgp\for estimation	Logistic P1	Logistic P2	Logistic P3	Probit P1	Probit P2	Probit P3
Logistic P1	0.362	**0.335**	0.845	0.343	0.357	0.747
Logistic P2	0.469	**0.386**	0.757	0.427	0.420	0.849
Logistic P3	0.878	**0.742**	0.803	0.794	0.806	1.002
Probit P1	0.389	**0.271**	0.608	0.275	0.390	0.809
Probit P2	0.339	0.328	0.829	0.405	**0.283**	0.888
Probit P3	1.025	**0.649**	0.919	0.825	0.702	0.837

In the following table 2(B) we give the results of correlation of the ranks estimated by all these models. In this case again Logistic P2 model is best in five cases showing highest correlation coefficients as compared to other models. However when we use Probit P2 model for data generating process then the same model gives the best result in re-estimation of ranks. Finally we can conclude that in overall situation IRT Logistic P2 model is best choice for estimation of true abilities.

Table: 2(B) Comparison between different IRT models on the basis of correlation between estimated ranks and true ranks.

Models for dgp\for estimation	Logistic P1	Logistic P2	Logistic P3	Probit P1	Probit P2	Probit P3
Logistic P1	0.908	**0.921**	0.484	0.917	0.907	0.502
Logistic P2	0.827	**0.871**	0.424	0.847	0.866	0.244
Logistic P3	0.372	**0.551**	0.327	0.530	0.498	0.086
Probit P1	0.786	**0.944**	0.699	0.941	0.753	0.521
Probit P2	0.917	0.884	0.426	0.879	**0.935**	0.359
Probit P3	0.203	**0.644**	0.078	0.471	0.572	0.284

5.3. Real life examples.

In this section we present the results of two different studies.

In study1, we estimate the abilities and ranks of the students from class 10^{th} belonging to different schools in Malakand district of NWFP. Abilities are estimated on basis of result of MCQ type test (consists of 50 questions). There were about 400 students who took parts in the test. Since we want to compare the estimated abilities of these students with their final board

examination result (we use this result as a proxy of true abilities), so we are required their final board examination results. Due non-availability of their final examination result and due to tedious nature of IRT analysis without a specialized software and manual marking, 80 students were selected randomly for final analysis.

In second study, we estimate the abilities of girl students who took admission in BS mathematics in IIU Islamabad on the basis of the results of entry test. We select only those students who have finally selected for admission. There were only 56 students and the entry test consists of 100 questions. We compare their estimated abilities with their final examination result of intermediate class.

For both the studies we apply Logistic P2 model which is the most power full model as we have proved in the last section. Mean absolute error of abilities estimated by IRT and CTT with the final year abilities (used as a proxy of true abilities) is given in the following table 3(A) for both studies. From table we can see that MAE of IRT abilities for both studies is smaller than CTT and there is 25% improvement in abilities estimation by IRT as compared to CTT for study 1 and for study 2 there is 6.7% improvement.

Table: 3 (A) Mean absolute error of the ability Indices measured by CTT and IRT with the final exam result

	Annual Board result's	
	Study 1	Study 2
CTT marks	0.111	0.262
IRT marks	0.088	0.246
Improvement by using IRT	25%	6.4%

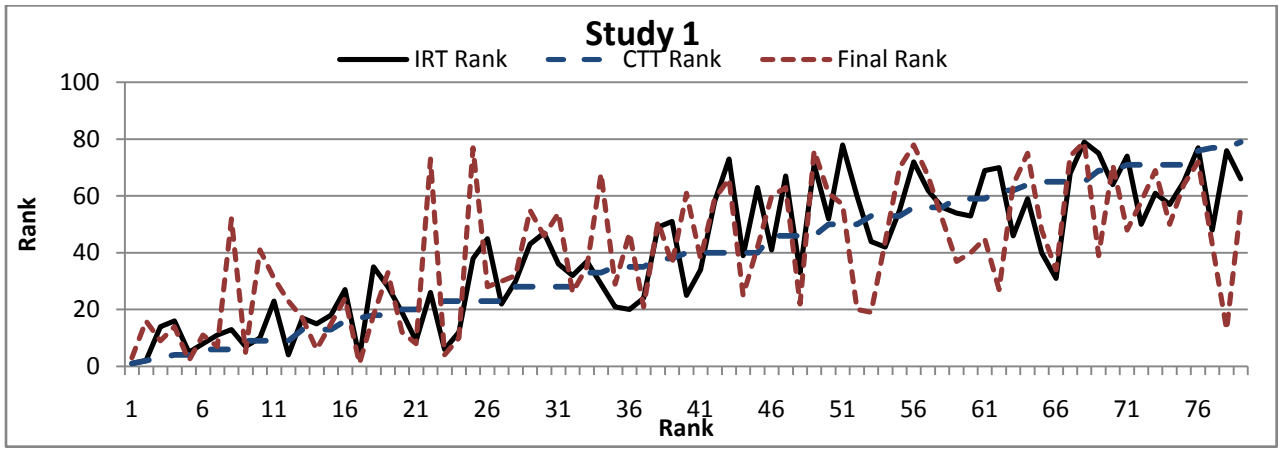
Ranks of the individuals, whose take part in the competition exam, play an important role in selection criteria for different fields, such as selection for job, for scholarship award etc. In CTT there is a big chance that two or more students having same scores and gets same ranks. However IRT solve this problem because in IRT score depends upon some assumptions, like difficulties of the parameters, discriminations of the parameters etc. In this study we estimate the ranks of the students under both CTT and IRT approach for both studies and compare them with final ranks. From the table 4(B), for study 1, we can note that correlation of IRT ranks with final ranks is 68% while CTT correlation is 59%. There is 15% improvement by estimating ranks under IRT. While for study 2 there is 6.5% improvement. In both the studies IRT estimation

technique is superior to CTT. From the Graph 2 given below, we see that IRT ranks are more correlated than CTT ranks. The table of ranks for both studies is given in appendix B. From the table of ranks, for study 1, we can see that the student having true rank 1 is at IRT rank 3 and CTT rank 17. Also we can see that there are 2 students having same rank 4 in CTT while in IRT there is no one case where two or more students having same ranks. Similar situation is for study 2.

Table: 3(B) Correlation between ranks of individuals measured by CTT, IRT and final exam's ranks

	Annual Board result's Rank	
	Study 1	Study 2
CTT Rank	.59	.32
IRT Rank	.68	.34
Improvement by using IRT	15%	6.5%

Graph 2: Graph of ranks of Individuals under both approaches along with true (Final) ranks.



6. Conclusion:

For small sample size 100, this Monte Carlo study indicates that CTT-based and IRT-based examinee ability estimates are very comparable and highly correlated (.95 for Probit P2 model) with each other, indicating that estimation of the individual's ability indices with different measurement theories will lead to the similar results. This is in accordance with the findings of Courville (2004), Fan (1998), Stage (1998), and MacDonald and Paunonen (2002). This Monte

Carlo studies proved that estimates of true abilities under Item Response Theory is more reliable than Classical methods. Since there are about six different models which have been used in IRT, in this regard this Monte Carlo study also provides a valuable finding that Logistic P2 model is best choice for IRT ability indices estimation. We apply this estimation technique with more power full model on real life examples and compare the estimates of true abilities of the examinee with their final board examination results (used as a proxy of true ability) and concluded that again IRT estimates of abilities and ranks of the examinee are more correlated with final exam's results as compared to CTT results. We also found that in both real life examples, the scores from both CTT and IRT are about 90% (on average) correlated with each other which supports Courville (2004, p.113) assertion that when scores obtained by two approaches are correlated they correlate by degree of 0.90 or higher; thus it is really hair splitting to argue about any difference between the two approaches

The result of this study support the claim of (A. Hotui, 2006) that MCQ type test can be used to measure high order skills because it was observed that the item which were falling in domain of higher cognitive skill were difficult and thus those students who failed to give correct response got more penalty in terms of losing score (Zaman & Atiq, 2008).

As the efficient allocation of resources is heart of economic problems, measuring abilities with improved accuracy can be helpful in this regard. This study help full for the selection of more capable persons in different fields such as for scholarship awards, for job in some reputable institutions, and for admission in educational institutions etc. By using traditional method of selection we can select less capable candidate which leads us to allocate resources to less efficient way.

7. References:

Baker, F. B. (1992). Item response theory: Parameter estimation techniques. New York:

Marcel Dekker.

Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika , 46*, 443-459.

Bock, R., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika , 35*, 179-197.

Cantrell, C.E. (1999). Item response theory: Understanding the one-parameter rasch model. In B. Thompson, *Advances in social science methodology* (Vol. 5, pp. 171-192). Stamford, CT: JAI Press.

Courville, T. (2004). *An Empirical Comparison of Item Response Theory And Classical Test Theory Item/Person Statistics: PhD thesis.* A & M University: Texas.

Hambleton R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston, MA: Kluwer Academic Publishers.

Hambleton R. K., Swaminathan H. & Rogers H. J. (1991). Fundamentals of item response theory. *Newbury Park,* CA: Sage.

Hambleton R. K. and Russel W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice , 12*, 38-47.

Henson, R. (1999). *Understanding the one-parameter Rasch model of item response theory*. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED 428 078).

Jimelo L. Silvestre-Tipay. (2009). Item Response Theory and Classical Test Theory: An Empirical Comparison of Item/Person Statistics in a Biological Science Test. The International Journal of Educational and Psychological Assessment. Vol. 1, Issue 1

Lawley, D. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh , 61A*, 273-287.

Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: theory and practice.* Dordrecht: Kluwer Academic Publisher.

Lord, F. (1952). A theory of test scores. *Psychometric Monograph* (7), Psychometric Society.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum Associates.

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

MacDonald P. and Paunonen S. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement , 62*, 921-943.

Richard J. Patz and Brian W. Junker (1999). Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses. Journal of Educational and Behavioral Statistics, Vol. 24, No. 4

Robert K. Tsutakawa and Michael J. Soltys (1988). Approximation for Bayesian Ability Estimation. Journal of Educational Statistics, Vol. 13, No. 2, pp. 117-130

Sheng, Y. (2005). *Bayesian Analysis of Hierarchical IRT Models: Comparing and Combining the Unidimensional & Multi-unidimensional IRT Models" PhD Thesis.* University of Missouri: Columbia.

Sheng, Y. (2008). Markov Chain Monte Carlo Estimation of Normal Ogive IRT Models in MATLAB. *Journal of Statistical Software , 25* (8).

Sotaridona, L. S., Pornel, J. B., & Vallejo, A. (2003). Some Applications of Item Response Theory to Testing. *The Philippine Statistician , 52* (1-4), 81-92.

Stage, C. (1999). *A comparison between item analysis based on item response theory and classical test theory: A study of the SweSAT test READ.* (Educational Measurement No31). Umea University: Department of Educational Measurement.

Terry A. Ackerman. (1991). Reviewed work(s): Item Response Theory: Parameter Estimation Techniques. by Frank B. Baker Source: Journal of the American Statistical Association, Vol. 88, No. 422.

Thomas, D. R., & Andre Cyr. (2002). *Applying Item Response Theory Methods to Complex Survey Data.* SSC Annual Meeting: Proceedings of the Survey Methods Section.

Troy Gerard Courville. (2004). An empirical comparison of Item Response Theory and Classical Test Theory item/person statistics. Ph.D thesis. Submitted to the Office of Graduate Studies of Texas A&M University

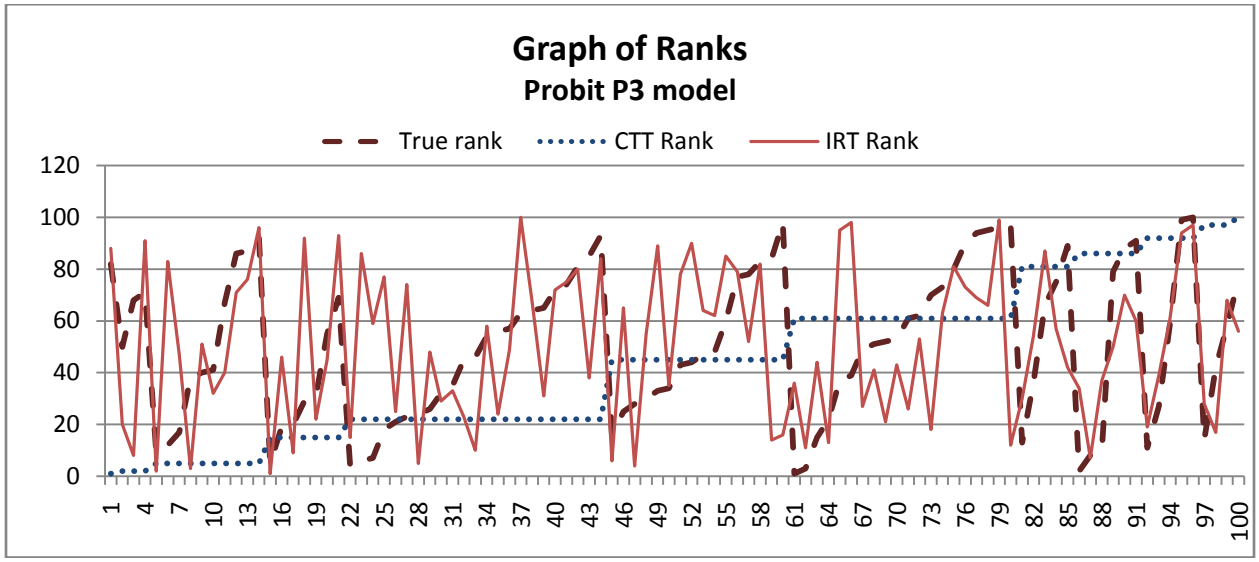
Xitao Fan. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educational and Psychological Measurem. Vol.58, No.3

Wiberg, M. (2004). *Classical Test Theory vs Item Response Theory: An evaluation of the theory test in the Swedish driving-license test.* Working paper: EM No 50, UMEA University.

Zaman, A., & Atiq-ur-Rehman. (2008). *Students Ranking, based on their Abilities on Objective type test: Comparison of CTT and IRT.* Accepted in the Confrence, "teaching and learning" Rothenburg, Germany

Appendix A:

Graph 1: Graph of ranks of individuals estimated by Probit P3 model with artificial data.



Appendix B:

Ranks estimated under both CTT and IRT along with true final ranks for both studies.

Study 1			Study 2		
IRT Rank	CTT Rank	Final Rank	IRT Rank	CTT Rank	Final Rank
1	1	3	1	1	4
2	2	16	2	2	24
3	17	1	3	3	5
4	9	23	4	4	8
5	4	2	5	4	23
6	23	4	6	8	11
7	9	5	7	8	25
8	6	11	8	4	41
9	20	8	9	11	14
10	9	41	10	8	6
11	6	7	11	13	19
12	23	10	12	4	16
13	6	52	13	11	35
14	3	9	14	14	25
15	13	6	15	17	34
16	4	14	16	17	17
17	13	17	17	14	3
18	13	15	18	17	11
19	20	12	19	17	54
20	35	47	20	29	43
21	35	29	21	17	25
22	23	30	22	24	43

23	9	31	23	22	11
24	35	21	24	29	35
25	40	61	25	34	19
26	22	73	26	34	35
27	16	24	27	22	40
28	18	33	28	24	55
29	33	68	29	34	2
30	28	32	30	24	51
31	65	34	31	14	39
32	28	26	32	24	41
33	46	22	33	29	1
34	40	38	34	29	47
35	18	18	35	42	49
36	28	54	36	34	9
37	33	35	37	42	46
38	23	77	38	29	32
39	40	25	39	45	18
40	65	48	40	34	19
41	46	60	41	24	56
42	53	44	42	51	51
43	28	55	43	34	6
44	53	19	44	45	35
45	23	28	45	34	9
46	62	64	46	34	45
47	28	46	47	45	30
48	77	43	48	45	51
49	38	51	49	44	15
50	71	58	50	55	19
51	38	36	51	45	31
52	50	61	52	55	50
53	59	40	53	45	32
54	59	37	54	51	47
55	53	70	55	53	28
56	56	52	56	53	28
57	71	50			
58	40	59			
59	64	75			
60	50	20			
61	71	69			
62	56	67			
63	40	42			
64	69	71			
65	71	64			
66	79	56			
67	46	63			
68	65	74			

69	59	45			
70	62	27			
71	46	76			
72	56	78			
73	40	66			
74	71	48			
75	69	39			
76	77	13			
77	76	72			
78	50	57			
79	65	79			