PROBABILITY THEORY AND APPLICATION OF ITEM RESPONSE THEORY

Zairul Nor Deana Md Desa, [1]Adibah Abdul Latif
Department of Foundation Education
Faculty of Education
Universiti Teknologi Malaysia
81310 UTM Skudai
Johor, Malaysia
zairul@utm.my
[1] p-adibah@utm.my

ABSTRACT

Since 1970s item response theory became the dominant area for study by measurement specialist in education industry. The common models and procedures for constructing test and interpreting test scores served the measurement specialist and other test users well for a long time. In this paper, a number of assumptions about the classical test theory are discussed. In addition, the contemporary test theory that is item response theory (IRT) will be discussed in general. There are four strong assumptions about the probability theory of item response theory such as the dimensionality of the latent space, local independence theory, the item characteristics curves and the speededness of the test. Furthermore, this discussion aimed to assist departments of education in considering the IRT that contributed in introducing the use of computerized adaptive testing (CAT) as they move to transition testing programs to online in the future.

Keywords: Item Response Theory; Latent Trait; Classical Test Theory; Computerized Adaptive Testing; Probability Theory.

INTRODUCTION

Educational assessment is very important in education process compared to the evaluation and judgment. Currently, the scenario of the education in Malaysia is emphasized on how the students and teachers can help each other to improve the ability and performance rather than the achievement in examination. According to Bloom in Payne (2003), assessment is an analysis process in individual life towards the criterion and the environment that related to the individual lives. It is included on how the individual face with all the stresses, problems and crisis using his / her own abilities and strengths. Therefore, the standardized test must be designed effectively to measure student's achievement.

One of the solutions in constructing good test that can measure the ability precisely is by using the item response theory (IRT) in building up the test. IRT models are mathematical functions that specify the probability of a discrete outcome, such as a correct response to an item, in terms of person and item parameter. Person parameters may represent the ability of a student or the strength of a person's attitude. Items may be questions that have incorrect and correct responses or statements on questionnaires that allow respondents to indicate level of agreement.

Before a user tends to use the IRT, he or she has to assure that the dataset met the assumptions of IRT to ensure that they choose the correct model. If not the model is considerably poor and ultimately questionable. In this paper, we will discuss four assumptions of item response models that are 1) the dimensionality of the latent space, 2) local independence, 3) the item characteristic curves, and 4) the speededness of a test which promote for better understanding on IRT models.

DIMENSIONALITY OF THE LATENT SPACE

The dimensionality of item response models is defined on how many abilities were tested to the examinees in a test. The examinee's ability is called examinee's latent traits due to its unobservable value. If item response models assume only a single ability or a homogeneous set of item to be tested then the dimension of items in a test are referred as *unidimensional*

such as might be found on vocabulary test. Certainly, there are a lot of factors that can affect examinee's performance such as their cognitive level, personality, level of examinee's motivation, ability to work quickly, knowledge of the correct use of answer tools, etc. These dominant factors are referred to as the abilities measured by the test [1]. For each test, the unidimensionality assumption should be checked so that the administered test measures the same trait at each time. Moreover, if an examinee's performance was tested by more than one latent variable then the IRT is referred to as *multidimensional* models (MIRT). So, it is assumed that there are *k* latent traits which define a *k*-dimensional latent space, and the examinee's position for each trait is determined by its location in the latent space. For example, a test consists of both biology and chemistry items are probably not sufficient to be considered as homogeneous IRT.

The placement of the examinee and the locations of the item are linearly related in measuring the examinee's performance. The illustration of this generic relationship has been shown by Boeck and Wilson (2004) as in Figure 1.

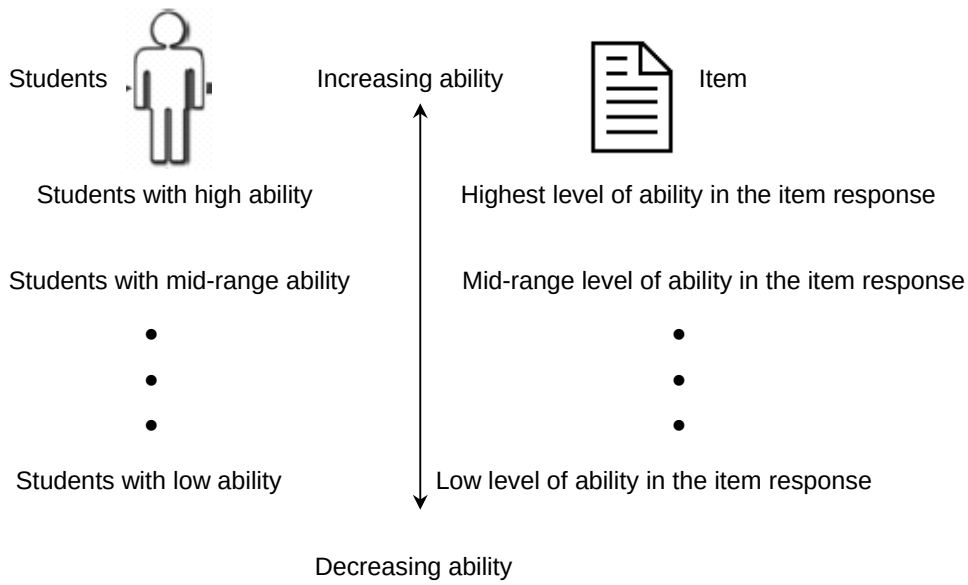

Figure 1: A generic illustration student-ability flow

Let say we have a test with *n* items was administered for a population of examinees. There are *r* subpopulation of examinees and the ability or latent traits is denoted as θ_r. The linear relationship between the performance of the examinee and the ability can be shown in the regression line as in Figure 2. At each ability level, the distribution of the test scores is conditional distribution to an ability level. If we have three groups of examinees (A,B and C), the unidimensionality of the test shows that the conditional distributions of the test score for these three groups would be equal. Contrary, at a given ability, if the test gives the examinees' performance differently then the test is said to be multidimensional. The relationship between the performance of the examinee and the ability of multidimensional test can be shown in the regression line as in Figure 3.

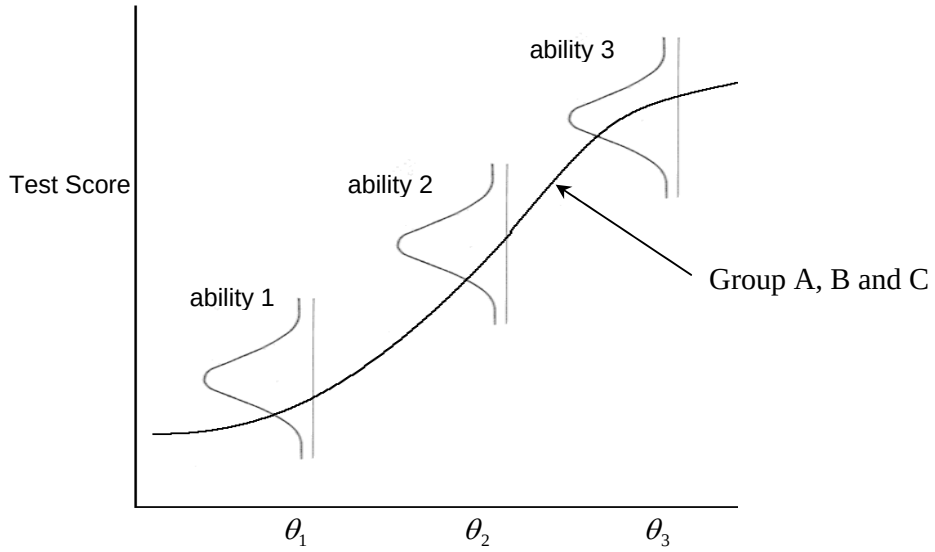

Figure 2: Unidimensional of conditional distribution of test scores at three ability levels

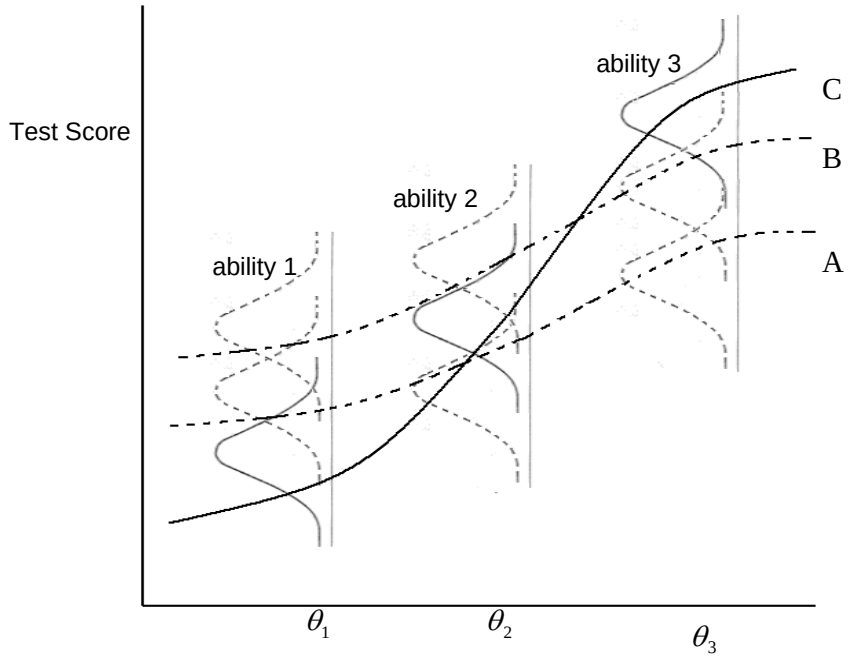

Figure 3: Multidimensional of conditional distribution of test scores at three ability levels

In computer adaptive test (CAT), the dimensionality of the test is significantly important. A CAT requires the examinee to response to a selected test items rather than to a standard test form. If the item pool consists of unidimensional items, then a CAT is set to a parallel test in random that promotes for equivalent measurement (Wainer, 1990, p. 211). This assumption affects the construct validity of a test. This assumption has to be checked to ensure that the constructed item for a test item pool measure a single latent trait. To check this, Hambleton and Swaminathan have recommended using the Kuder-Richardson Formula 20 (KR-20) to address the dimensionality of a set of test items. The KR-20 should be used if we know the test length and the heterogeneity of the examinee. In addition, they said that the unidimensionality of a set of test item can be checked using tetrachoric correlation or phi

correlation in factor analysis. However, these checking procedures required complex formulation with some numerical integration and the decision about the eigenvalues of the interitem tetrachoric correlation matrix somewhat difficult to determine (Ackerman, 1996; Hambleton & Swaminathan, 1985, and Boeck & Wilson, 2004).

LOCAL INDEPENDENCE

In applying the IRT models, the so-called of *local independence* assumption is one of the important features. This assumption meaning that for every examinee's response $\mathbf{y}_{pi} = 0 \text{ or } 1$ (where 0 denotes an incorrect response and 1 denotes a correct response) to the items *i* are statistically independent. In other words, the examinee's performance on one item does not influenced by the correctness of answering the other items. When the IRT models met this assumption, the probability of that examinees of ability θ correctly answer item 1 and item 2 equals to the product of the probability of correctly answering item 1 and the probability of correctly answering item 2 that is:

$$P\left[Y_1 = y_1, Y_2 = y_2 \middle| \theta\right] = P\left[Y_1 = y_1 \middle| \theta\right] \cdot P\left[Y_2 = y_2 \middle| \theta\right]$$

generally the assumption of local independence implies the joint distribution as in Equation [1]

$$\begin{aligned} P\left[Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n \middle| \theta\right] &= P\left[Y_1 = y_1 \middle| \theta\right] \cdot P\left[Y_2 = y_2 \middle| \theta\right] \cdots P\left[Y_n = y_n \middle| \theta\right] \\ &= P_1[\theta]^{y_1} Q_1[\theta]^{1-y_1} P_2[\theta]^{y_2} Q_2[\theta]^{1-y_2} \cdots P_n[\theta]^{y_n} Q_n[\theta]^{1-y_n} \\ &= \prod_{i=1}^{n} P_i[\theta]^{y_i} Q_i[\theta]^{1-y_i} = L\left[\mathbf{y} \middle| \theta\right] \end{aligned} \qquad [1]$$

where $Q_i(\theta) = 1 - P(\theta)_i$. For example the probability of the dichotomous response of $\mathbf{y}_{pi} = (0\ 1\ 0\ 1\ 1)$ is equal to $Q_1 \cdot P_2 \cdot Q_3 \cdot P_4 \cdot P_5$ where P_i is the probability of the correct response and $Q_i = 1 - P_i$ is the probability of incorrect response. Consider that if the examinee of ability $\theta = 1.5$ have a 0.30 probability of answering item 1 correctly and a 0.50 probability of answering item 2 correctly, under local independence assumption the probability of the examinee answering both item correctly is $0.30(0.50) = 0.15$.

Note that not all cases hold the requirement of local independence. For instance, in the case where some examinees may have higher expected test score than other examinees with respect to the same ability (Hambleton & Swaminathan, 1985) and if the test item is long and the examinees learn while answering items. Accordingly, the items associated with one stimulus are likely to be more related to one another than to items associated with another stimulus (Kolen & Brennan, 2004). For example, when test are compose of sets of items that are based on common stimuli, such as items on reading passages or diagrams, then Equation [1] is unlikely to hold. The violated relationship between the assumption of local independence and of the unidimensionality, however, for both might hold closely enough for IRT to be used in many practical situations such as in CAT. If unidimensionality is met, then the local independence assumption is also usually met (Hambleton et al., 1991). However, local independence can be met even when the unidimensionality assumption is not (Scherbaum et. Al. 2006, Kolen & Brennan 2004).

ITEM CHARACTERISTIC CURVES

The *item characteristic curves* (or ICC – they are also referred to as the item response function; IRF, or item characteristics function or trace lines) show the connection between the means of the conditional probability as in Equation [1] with the regression of the item score on ability. The frequency of test scores of examinees of fixed ability is given by

$$f(x|\theta) = \sum_{\sum y_i = x} \prod_{i=1}^{n} P_i[\theta]^{y_i} Q_i[\theta]^{1-y_i}$$

[2]

where x is an examinee's test score and $x \in [0, n]$. The ICC for item i denoted by $P_i(\theta)$ as the probability of an examinee answering item i correctly with ability θ. For example, 35% of the examinees with ability $\theta = 1.5$ are expected to answer item 1 correctly with the probability is $P_1(1.5) = 0.35$.

There are no fixed mathematical function of ICC. If we considered only one latent ability, then the regression is referred to as an ICC. Otherwise such as in multidimensional models, the regression has been referred to as the item characteristic function. This can be explained by the conditional distribution as in Equation [1] is identical across different populations, so that IRT models typically assume a specified functional form for the item characteristic curves.

Universally, there are three mathematical models for the ICC to illustrate the relation of the probability of correct response to ability. Each model represents one or more parameters and the standard form of the model of the ICC is the cumulative form of the *logistic function*. First model are known as 1-PL or one-parameter logistic model and also known as *Rasch Model* that is given by

$$P_i(\theta) = \frac{1}{1 + e^{-(\theta_i - b_i)}}$$

[3]

where e is the constant 2.718, and b_i is the item difficulty parameter for item *i*. In other words, we can say that the proportion of items that a particular examinee with ability θ can answer item i correctly is given by Equation [3]. Figure 4(a) shows the "S" shaped of 1-PL model for three different levels of difficulty. The higher the value of b_i, the probability of correct responses increases as well. The typical values of b_i have the range $-3 \le b \le +3$ (Baker, 2001).

In application of IRT 1-PL model does not get a good fit to the data since the items are not always parallel (Wainer & Mislevy, 1990). Therefore, there are alternative models should be used to encounter this problem. We can either delete items that the curves show the slopes that are divergent or generalize the model to allow different slopes. In the *two-parameter logistic model*, ICCs vary in both slope and difficulty (some items are more difficult than others). In this model, there is additional parameter for each item. This parameter denoted as a and is often called as the item's discrimination. The 2-PL is

$$P_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

[4]

where a_i is the discrimination level of item i. The usual range of a seen in practice is $-2.80 \le a \le +2.80$ (Baker, 2001). To illustrate how the two-parameter model is used to compute the points on an ICC, consider the following example problem.

Let say $b = 1.0,\ a = 0.5$, the ability $\theta = 3.0$, and the logit of Equation [4] is as follows

$$\begin{aligned} L = LogP_i(\theta) &= Log1 - \log\left(1 - e^{-a_i(\theta - b_i)}\right) \\ &= Log1 - \log 1 + a_i(\theta - b_i)\log e \\ &= a_i(\theta - b_i) \end{aligned} \qquad [5]$$

substitute $b = 1.0,\ a = 0.5$, the ability $\theta = 3.0$ into Equation [5] as follows

$$L = a_i(\theta - b_i) = 0.5(3.0 - 1.0) = 1.0$$

Therefore, the value of $P_i(\theta)$ is

$$\begin{aligned} P_i(3.0) &= \frac{1}{1 + e^{-L}} \\ &= 1/1 + e^{-1.0} = 0.7311 \end{aligned}$$

Thus at an ability level of 3.0, the probability of answering correctly to this item is 0.7311. Table 1 shows the calculation of $P_i(\theta)$ for this item at seven ability level that ranges from +3.0 to -3.0. Figure 4(b) shows the typical ICC for the 2-PL model for the same difficulty levels.

Table 1: Item characteristic curve calculations under a two-parameter model, $b = 1.0,\ a = 0.5$

Ability, θ	***L***	$P_i(\theta)$
3	1	0.7311
2	0.5	0.6225
1	0	0.5000
0	-0.5	0.3775
-1	-1	0.2689
-2	-1.5	0.1824
-3	-2	0.1192

In general ability test, placement test or any types of test the multiple-choice item remains popular and therefore the facts in testing that the examinees will get item correct by guessing is possible. Neither 1-PL nor 2-PL models took the guessing phenomenon into consideration. It is possible that the examinees will guess the answer for difficult item correctly or he or she answered using skill or knowledge based other than the one we thought we were testing. This phenomenon can be corrected by modified the 2-PL models to include a parameter that represents the contribution of guessing to the probability of correct response. This modification has been first recommended by Allan Birnbaum in 1968 (Baker, 2001, Wainer & Mislevy, 1990). It adds a third parameter, *c,* that is a binomial form of the guessing parameter. This modified model called the three-parameter model or 3-PL models is shown in Equation [6] as follows

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

[6]

where c_i is the guessing parameter for item *i* or also known as the probability of getting the item *i* correctly by guessing alone. The range for c_i between zero and 1 but the practical value is $0 \le c_i \le 0.35$. Table 2 shows the calculation of in 3-PL models. The corresponding item characteristic curve is shown in Figure 4(c).

Table 2: Item characteristic curve calculations under a three-parameter model,

$b = 1.5,\ a = 1.3$ and $c = 0.25$

Ability, θ	***L***	$P_i(\theta)$
3	1.95	0.9066
2	0.65	0.7428
1	-0.65	0.5072
0	-1.95	0.3434
-1	-3.25	0.2780
-2	-4.55	0.2578
-3	-5.85	0.2522

It is common for IRT user to define the item response function before beginning their work. The 3-PL is the IRT model that is most commonly applied in large-scale testing applications.

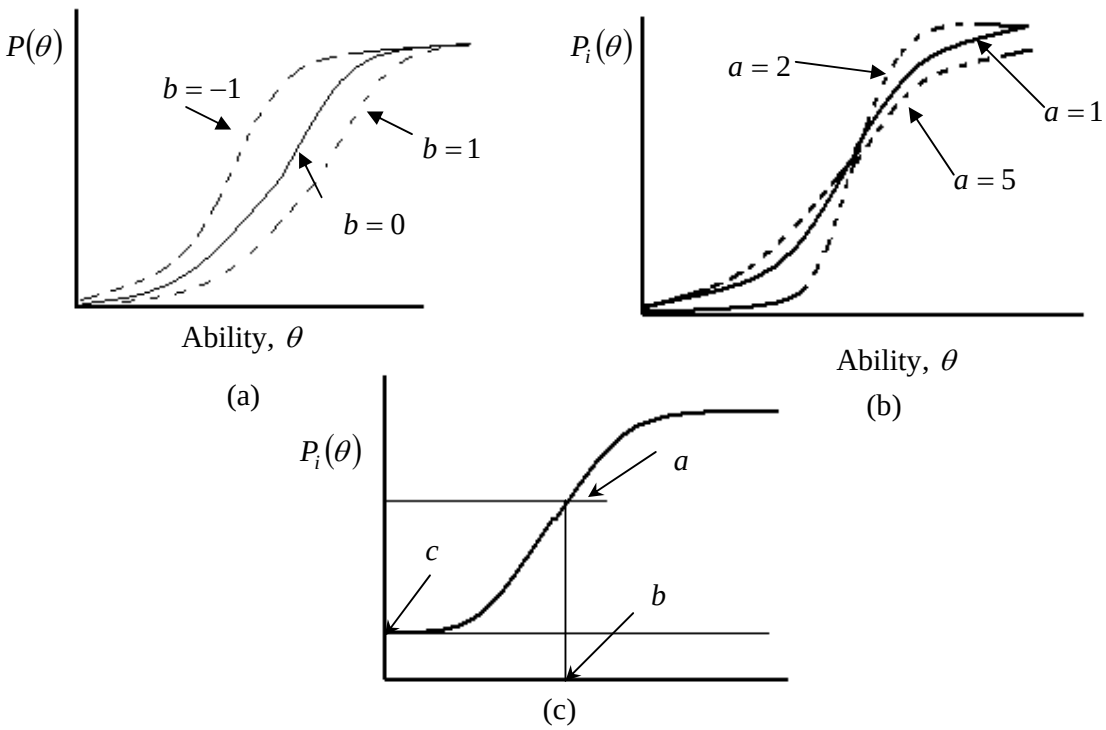

Figure 4: (a) Typical item characteristic curves for 1-PL models at three levels of difficulty; (b) typical item characteristic curves for 2-PL models; and (c) typical item characteristic curves for 3-PL models

SPEEDEDNESS OF THE TEST

Not many IRT users give their attention on the speediness of the test that may influence the examinee's performance. The failure of examinees to complete the test in the given time limit so that this factor does not contaminate ability scores estimates (Hambleton & Swaminathan, 1985). When speed becomes one of the factors, then in the IRT models are consists of at lest two traits affecting the examinees' performance: the latent traits and the time limit. The speediness of the test can be checked by identifying the number of examinees who fail to finish a set of test and the number of items they fail to complete. Donlon (1978) (in Hambleton & Swaminathan, 1985) provided the estimate of correlation between scores obtained under power and speed conditions for identifying the speediness of tests as follows

$$\rho\left(T_p, T_s\right) = \frac{\rho\left(Y_p, Y_s\right)}{\sqrt{\rho\left(Y_p, Y_p'\right)}\sqrt{\rho\left(Y_s, Y_s'\right)}}$$

and if we have administered parallel form of the test, then the speediness of the test can be computed using the speediness index as follows

$$\text{Speediness Index } = 1 - \rho^2\left(T_s, T_p\right)$$

Oshima (1994) discussed three factors to be considered in relation to speediness: proportion of the test not reached 5%, 10%, and 15%, response to not reached "blank and random response", and item ordering from random to easy and hard. He found that ability estimation was least affected by the speededness of the test in terms of correlation between true and estimated ability parameters.

APPLICATION OF ITEM RESPONSE THEORY

IRT is the body of the development of modern psychometric fields. The theory and technique of IRT for examining psychometric properties of measures are much more complex than the classical methods. However, there are a lot of advantages when applying the IRT against the classical test theory. IRT models are particularly important and as the foundation for adaptive testing (mainly in computerized adaptive testing, CAT). When a test has been administered via the computer, the computer can update the items selection due to the present examinee's performance at an ability. To match test items to the ability levels requires a large pool of items. With the right item bank and a high examinee ability variance, CAT can be much more efficient than a traditional paper-and-pencil test (P&P). In adaptive testing, the strategy for selection of item from items pool follows the following decision rules:If an examinee answer an item correctly, the next item should be more difficult. If an examinee answer incorrectly, then next item should be easier.

There are many advantages of applying IRT in computerized adaptive testing such as:

i. When the assumptions of the IRT models are actually met, IRT provides correspondingly stronger findings due to invariant parameters estimation. In contrast, the classical test theory estimates the parameters such as the item difficulty, item discrimination, and reliability in a specific way. The error scores are also assumed to be constant.
ii. IRT provides several improvements in scaling items and people
iii. The parameters of IRT models are generally not sample- or test-dependent. Thus, IRT provides significantly greater flexibility in situations where different samples or test forms are used.
iv. Items in adaptive testing of IRT models are chosen on present of examinee's ability. Therefore, the raw score is based on a weighted sum of the item responses. Contrary, in P&P, responses are equally weighted.
v. Test administered on computer will give different experience for examinees than taking a P&P test. Some of these differences are ease of reading passage or reviewing and changing answers, the effects of time limits, clearness of figures or diagrams and responding on a keyboard vs. responding on an answer sheet.

SUMMARY

IRT models are now an established approach of development of many testing method in educational measurement and evaluation. The purpose of this writing is to provide an introductory discussion of the main assumption, which has to be met, of many item response models. IRT models have been widely used due to its efficiency (fewer administered items) and control of correctness when given adequate items, every person can be measured with the same degree of correctness.

It is our belief that these techniques will lead to a greater understanding of the testing and measurement in education. As in our National Education Blueprint 2006-2010, the

government enhancing the alternatives assessment methods that endorse for better learning environment among students. Therefore, the contributions of IRT models should be consider among psychometricians and we have, seriously, to start talking (and applying) IRT models in our assessment system.

REFERENCE

Ackerman, T.(1996). Graphical Representation of Multidimensional Item Response Theory Analyses Applied Psychological Measurement. Vol. 20, No. 4, December 1996, pp. 311-329.

Baker, F. B. (2001). The Basics of Item Response Theory.2nd ed. USA: ERIC Clearinghouse on Assessment and Evaluation

De Boeck, P. & M. Wilson. (2004). Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. USA: Springer-Verlag.

Hambleton, R.K. & H. Swaminathan, (1985). Item Response Theory: Principles and Application. UK; Kluwer Nijhoff Pub.

Jinming Zhang & Ting Lu.(2007). *ResearchReport* : Refinement of a Bias-Correction Procedure for the Weighted Likelihood Estimator of Ability. Princeton, NJ: Educational Testing Services.

Kolen, M.J. & R. L., Brennan.(2004). Test Equating, Scaling and Linking: Methods and Practice. 2nd Ed. USA: Springer-Verlag.

Lihua Yao and Schwarz, R. D.(2006). A Multidimensional Partial Credit Model With Associated Item and Test Statistics: An Application to Mixed-Format Tests. Applied Psychological Measurement, Vol. 30 No. 6, November 2006, 469–492

Oshima, T.C. (1994). The Effect of Speededness on Parameter Estimation in Item Response Theory. Journal of Educational Measurement*,* Vol. 31, No. 3 (Autumn, 1994), pp. 200-219.

Principles, characteristics, and assessment, with an illustrative example Journal of Business Research 57 (2004) 184– 208.

Scherbaum, C.A., F. Scott, K. Barden, T. Kevin. (2006). Applications of item response theory to measurement issues in leadership research The Leadership Quarterly 17 (2006) 366–386.

Sitjsma, K. & Brian, W. J.(2006). Item response theory:past performance, present development and future expectation. Behaviormetrika, vol 33, No. 1, pg 75-102.

Tamás Antal. 2007. *ResearchReport* ; On Multidimensional Item Response Theory: A Coordinate-Free Approach. ETS, Princeton, NJ.

Wainer, H. (1990). Computerized Adaptive Testing; A Primer. New Jersey: Lawrence Erlbaum Associates.