Theoretical and Practical Comparison of Classical Test Theory and Item-Response Theory

Gholam Reza Kiany [*]
Associate professor of Applied Linguistics, English Department, Faculty of Literature & Humanities, Tarbiat Modarress University, Tehran, Iran
&
Sara Jalali
Urmia University & PhD Candidat at Tarbiat Modares University

Abstract
Classical test theory and item response theory are widely perceived as representing two very different measurement frameworks. Few studies have empirically examined the similarities and differences in the parameters estimated using the two frameworks. The purpose of this study was to examine how item statistics (i.e. item difficulty and item discrimination) and person statistics (i.e. ability estimates) behave under the two measurement frameworks i.e. CTT and IRT. The researchers tried to compare the two models from both theoretical and practical perspectives. For this purpose, first, a theoretical comparison of the two models was carried out; then, a sample of 3000 testees taking part in the English language university entrance exam was used in order to compare the two models practically. The findings showed that person statistics from CTT were comparable with those from IRT for all three IRT models. Item difficulty indexes from CTT were comparable with those from all IRT models and especially from the one-parameter logistic (1PL) model. Item discrimination indexes from CTT were somewhat less comparable with those from IRT.

[*] *E-mail address:* kiany_gh@modares.ac.ir

Address: Tehran- Ale Ahmad and Chamran Crossroads- Tarbiat Modares University- Faculty of Humanities- English Department
Phone: 021- 82884664

Keywords: Item Response Theory (IRT); Classical Test Theory (CTT); Person statistics; Item statistics

Introduction

In the theory of measurement, there are two major measurement frameworks: classical test theory (CTT) and item response theory (IRT). Differences are most evident in the statistical analysis underlying each theory.

Classical test theory (CTT)

Classical test theory (CTT) is best suited for traditional testing situations, either in group or individual settings, in which all the members of a target population, e.g. persons seeking college admission, are administered the same or parallel sets of test items. CTT has a number of underlying assumptions (cf. Bachman, 1990):

1. In this model, an observed score on a test consists of two components: a true score that is the result of an individual's ability level and an error score that is the result of factors other than the ability being tested. This assumption can be depicted in this formula:

$x = x_t + x_e$

where

x = the observed score

x_t = the true score

x_e = the error score.

As it can be observed, the technical aspect of this assumption is additivity i.e. the true and error scores add to form the observed score. In other words, the observed score is assumed to be the sum of the true and error scores. Similarly, the variance of a set of test scores can be characterized as comprising two components: $s^2_x = s^2_t + s^2_e$

where

s^2_x = the observed score variance

s^2_t = the true score variance

s^2_e = the error score variance.

2. The second assumption is that error scores are unsystematic or random and are uncorrelated with true scores ($r_{te} = 0$). Therefore, according to the CTT model, measurement error is the variation in a set of test scores that is unsystematic and random.

CTT defines two sources of variance in a set of test scores: the true score variance, which is due to differences in the ability of the individuals tested, and measurement error, which is unsystematic or random.

3. Another assumption of CTT is the concept of parallel tests. According to CTT, two tests are parallel if, for every group of testees taking both tests: a) the true score on one test is equal to the true score on the other, and b) the error variance for the two tests are equal. In other words, parallel tests are two tests of the same ability that have the same means and variances and are equally correlated with other tests of the ability i.e. $\bar{x}_1 = \bar{x}_2$, $s^2_{x1} = s^2_{x2}$, and $r_{x1y} = r_{x2y}$

where

$\bar{x}_1$ and $\bar{x}_2$ = the mean scores of the two parallel tests

s^2_{x1} and s^2_{x2} = variances of the two parallel tests

r_{x1y} and r_{x2y} = the correlation between the scores from a third test 'y' and the tests x_1 and x_2 respectively, 'y' is any other test of the same ability.

4. The concept of reliability in CTT is described in the context of parallel tests. In parallel tests, the true score on one test is equal to the true score on the other test. The error scores of both tests are assumed to be random and will be uncorrelated. Because of the influence of the random error scores, the correlation between observed scores of parallel tests will be less than perfect. The smaller the influence of the error scores, the more highly the parallel tests will be correlated. If the observed scores on two parallel tests are highly correlated, this shows that influences of the error scores are minimal, and they can be considered reliable indicators of the ability being measured. From this comes the definition of reliability as the correlation between the

observed scores on two parallel tests, which is symbolized as r_{x1x2}. This definition provides the basis for all estimates of reliability within CTT.

CTT is not a complete model and has a number of shortcomings and problems:

1. Researchers often speak of the reliability of a given test; strictly speaking, reliability refers to the test scores and not the test itself. Since reliability is a function not only of the test, but also of the performance of the individuals who take the test, any given estimate of reliability based on CTT is limited to the sample of test scores upon which it is based.

2. CTT treats error variance as homogeneous in origin, consequently, different sources of error may be confused, or confounded with other and with true score variance. This is because it is not possible to examine more than one source of error at a time, although the test performance may be influenced by many different sources of error simultaneously.

3. CTT considers all errors to be random. However, there are some errors, which are systematic and happen regularly like cultural background, ethnicity, field-dependence, etc. These systematic errors most of the time result in test bias which is completely ignored in this model.

4. In CTT, the most important pieces of information are total scores or raw scores. Every testee is given one score which shows his performance on the whole exam. Items do not play any significant roles in this model.

5. The other problem with CTT is its "circular dependency" i.e. 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 situation" like CAT (Fan, 1998, p.357). Therefore, CTT statistics are sample dependent in that as the sample changes, the estimators would change, and consequently the estimators are not generalizable across populations.

The only information that is available for predicting a testee's performance on a given item is the index of difficulty or 'p' which is the proportion of individuals in a group that responded correctly to the item. "Thus, the only information available in predicting how an individual will answer an item is the average performance of a group on this item" (Bachman, 1990, p. 203).

It should be mentioned that the proportion correct (p) is dependent not only on the difficulty of the item itself but also on the ability of the testees who are used in calculating the value. This is known as sample dependence. In other words, with different sample of testees, the value could be different. Because of this, the sample upon which the statistic is calculated should be genuinely representative of the population of testees for whom the test is designed. Unless this is the case, score meaning is compromised. In addition, score meaning is compromised when the test is utilized for a purpose or population for whom it was not originally intended (Fulcher & Davidson, 2007).

"For those who view the raison d'être of measurement as permitting one to differentiate among examinees, the key indicator of an item's value is its discrimination index" (Millman & Greene, 1989, p. 359). The ability of an item to discriminate between higher ability testees and lower ability testees is known as item discrimination, which is expressed statistically as the Pearson product-moment correlation coefficient between the scores on the item (e.g. 0 and 1 on an item scored right-wrong) and the scores on the total test. When an item is dichotomously scored, this estimate is computed as a point-biserial correlation coefficient (Fan, 1998). In other words, when one dichotomous variable is to be correlated with a continuous variable point-biserial correlation coefficient is available. The most frequent occasion for using this formula is in correlating a dichotomous test item (e.g. pass-fail or right-wrong) with total scores on a test (Nunnally, 1993).

As Nunnally (1993) mentions, when one variable is dichotomous and the other continuous, the biserial correlation can be employed in place of the point-biserial correlation.

> *Item discrimination statistics focus not on how many people correctly answer an item, but on whether the correct people get the item right or wrong. In essence, the goal of an item discrimination statistics is to eliminate items that do not function as expected in the tested group. ...the index of discrimination can range from -1 to 1. A positive index indicates that a higher proportion of the upper group answered the item correctly, while a negative item discrimination index (D) indicates that a larger proportion of the lower group answered the item correctly (Courville, 2004, p. 38-39).*

Generally, items with an r_{pbi} of 0.25 or greater are considered acceptable, while items with a lower value would be rewritten or excluded from the test (Henning, 1987). "As with item difficulty, measures of discrimination are sensitive to the size of the sample used in the calculation, and the range of ability represented in the sample. If the sample used in the field trials of items is not large and representative, the statistics could be very misleading" (Fulcher & Davidson, 2007, p. 104).

> *If we are interested in the correlation between the variable that the item measures and the continuous criterion measure, and if we may assume that the thing measured by the item is continuously and normally distributed in the population, the biserial r is the coefficient we want. ...If we are interested in how well we can predict the criterion from the item or how much it can contribute to a total score, with its own score limited to 0 and 1, the point-biserial r is the coefficient to compute. The test theory that regards a total score as*

> *the summation of item scores assumes this type of correlation" (Guilford, 1954, p. 427).*

Item response theory (IRT)

"IRT, also known as latent trait theory, is model-based measurement in which trait level estimates depend on both persons' responses and on the properties of the items that were administered" (Embreston & Reise, 2000, p. 13).

Item response theory (IRT) provides more item, person, and test information than CTT. Here, item and response are both important. IRT is, for some researchers, the answer to the shortcomings of CTT. IRT is often referred to as 'latent trait theory', 'strong true score theory', or 'modern mental test theory'. It is a modeling technique that tries to describe the relationship between a testee's test performance and the latent trait underlying the performance. It provides a basis for estimates of measurement error that are not dependent upon particular samples of individuals, and for estimating differential measurement error at different ability levels. In other words, it is a new and different way of looking at the entire psychometric process, one that is much more mathematically and conceptually complex and requires a new and deeper level of thinking to appreciate. IRT focuses on items rather than overall test scores, it also helps how item parameters such as discrimination, difficulty and guessing parameter can be calculated.

An important characteristic of IRT is that it is parameter invariant. That is, the information provided by IRT regarding item parameters or item statistics (i.e. item difficulty and item discrimination), unlike that provided by CTT, is invariant to the sample used to generate the item and test information. This is because the mathematical model used to derive item parameters in IRT is derived based on the estimated latent trait (θ) and not the test taker's total score. Psychological constructs are conceptualized as latent traits. Latent traits are unobservable entities that influence observable variables such as test scores and item responses (Crocker & Algina, 1986). In fact, test score or item response gives

information on a testee's standing on the latent trait (θ). Information obtained from one sample using IRT, assuming it is sufficiently large but not necessarily representative of the target population will be equivalent to that obtained from another sample, regardless of the average ability level of the testees who took the two tests. The same is not true for CTT. Therefore, in contrast to the "circular dependency", IRT person statistic is item-free (i.e. would not change if different items were used) and the item statistics are person-free (i.e. would not change if different persons were used).

The IRT framework includes a group of models, and "the applicability of each model in a particular situation depends on the nature of the test items and the variability of different theoretical assumptions about the test items. For test items that are dichotomously scored, there are three IRT models" i.e. one-parameter IRT model, two-parameter, and three-parameter IRT models (Fan, 1998, p. 358).

Different IRT models can be characterized in terms of differences in their general form, and in the types of information, or parameters, about the characteristics of the item itself. The types of information about item characteristics may include (Bachman, 1990):

1. The degree to which the item discriminates among individuals of differing levels of ability (the 'discrimination' parameter 'a').
2. The level of difficulty of the item (the 'difficulty' parameter 'b').
3. The probability that an individual of low ability can answer the item correctly (the 'pseudo-chance' or 'guessing' parameter 'c').

Over the past twenty-nine years, since Lord's 1980's book, IRT has become the "jewel of large-scale test construction programs". As Fan (1998) mentions,

> *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 and response statistics. Theoretically, such*

> *relationships are not entirely clear, except that the two types of statistics should be monotonically related under certain conditions. But such relationships have rarely been empirically investigated, and, as a result, they are largely unknown (p. 360).*

CTT and IRT are widely perceived as representing two very different measurement frameworks. Few studies have empirically examined the similarities and differences in the parameters estimated using two frameworks.

The purpose of this study was to examine how item statistics (i.e. item difficulty and item discrimination) and person statistics (i.e. ability estimates) behave under the two competing measurement frameworks i.e. CTT and IRT.

Studies by Courville (2004), Fan (1998), Hwang (2002), Lawson (1991), MacDonald and Paunonen (2002), Skaggs and Lissitz (1986, 1988) and Stage (1998a, 1998b, 1999) have all referred to little difference between IRT and CTT estimates. In Stage's (1999) work with the SweSAT test READ, she states 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" (p. 19-20).

This study was significant in three ways: First, by providing a comprehensive description and comparison of CTT and IRT models, it tried to create a clear theoretical picture of the two models. After that, through a practical data analysis, the two models, and the differences and similarities between the two models were discussed and analyzed. Here, real data were utilized and the basic tenets of the two models i.e. person statistic and item statistics were compared with each other. Therefore, TEFL researchers would have the chance to deal with the two models practically and in a tangible way. On the other hand, very

few studies have compared CTT and IRT for item and person analysis. As Fan (1998) mentions,

> *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 an anomal (p. 361).*

As Hening (1987) mentions, "it may be necessary systematically to inform the public and legal system that it is more desirable to hold measurement error constant rather than to hold constant the number and exact examples of items encountered" (p. 137).

> *Unfortunately, the view that the argument is moot seems to have occurred largely in the vacuum of empirical evidence, because the literature fails to show that this important premise has been subjected to systematic and rigorous empirical investigation. It is my view that in psychological measurement, as in any other areas of science, theoretical models are important in guiding our research and practice. But the merits of a theoretical model should ultimately be validated through rigorous empirical scrutiny (Fan, 1998, p. 15).*

The research hypotheses are as follows:
Ho1. There is not any difference between the CTT-based and IRT-based person statistic (testee ability estimate) in the three IRT models.

Ho2. There is not any difference between the CTT-based and IRT-based item difficulty statistic (estimate) in the three IRT models.

Ho3. There is not any difference between the CTT-based and IRT-based item discrimination statistic (estimate) in the two IRT models.

Method

Subjects

The participants in this study were testees who took the English part of the foreign language university entrance exam in 2006. 3000 testees were randomly selected out of the whole population who took the exam in 2006. Their performances were analyzed regarding person and item statistics.

Instrumentation

The items analyzed in this study came from the English part of the foreign language university entrance exam which is a high-stakes test used for admissions to universities in Iran. The English part of the foreign language university entrance exam contains 70 multiple-choice items forming six subparts: structure (10 items), vocabulary (20 items), word order (5 items), language function (5 items), cloze test (15 items), and reading comprehension (15 items).

The BILOG software was used for carrying out the IRT analyses. The SPSS software was utilized for the CTT analyses.

Procedure

The answer sheets of 3000 participants were randomly selected out of a pool who took part in the foreign language university entrance exam in 2006, which was 6000. Their answer sheets were collected and all answers to all items were entered to the software for analysis. The necessary statistical tests and procedures were carried out in order to find person statistic (testee ability estimates) and item statistics (item difficulty and item discrimination). After that, the researcher compared the results of the CTT-based and IRT-based person and item statistics. Then, the similarities and differences between the two models were found in order to answer the research questions.

Results

Using IRT and obtaining dependable results are only possible when the first and foremost presupposition of IRT is met i.e. unidimansionality of the test. As mentioned before, unidimansionality states that the items in a test measure a single unidimensional ability or trait, and that the items form a unidimensional scale of measurement.

Since the test consisted of six subparts and each subpart tested a specific area, unidimensionality was checked for each subpart separately. The subparts consisted of structure (10 items), vocabulary (20 items) word order (5 items), language functions (5 items), cloze test (15 items), and finally reading comprehension (15 items).

In order to check unidimensionality, the data was analyzed by using the TESTFACT software. A factor analysis was carried out through this software and the eigen values were checked. It was found that a single dominant factor underlie the responses, therefore, the unidimensionality assumption was met. The following table shows the results. As it can be observed here, there was one major factor, which accounted for more than 17% of the variance of the scores in each subpart (Yen, 1985). This shows that every subpart had one major underlying factor. It should be mentioned that some subparts consisted of just 5 or 10 items; therefore, it would be acceptable if the percent of variances for these subparts were low. Based on these results, the researchers concluded that the unidimensionality assumption for the IRT models held for the data used in this study.

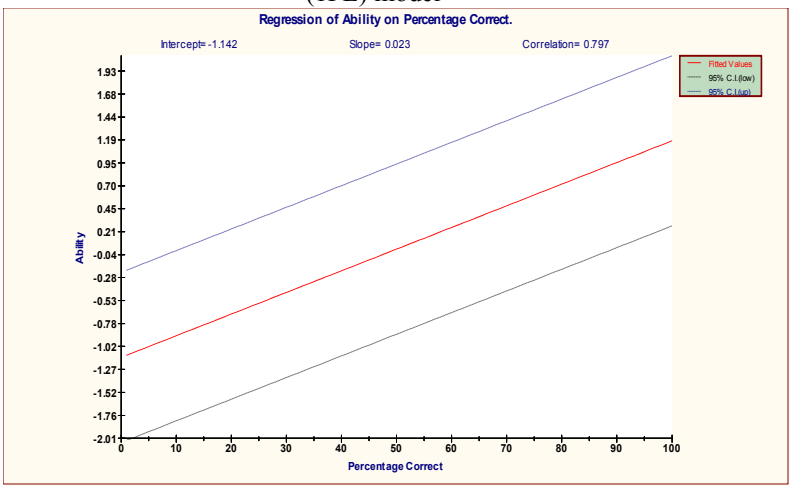
Table 1
Eigen value for all subparts

Subpart	Percent of variance
Structure	19.73
Vocabulary	19.20
Word order	17.16
Language function	22.93
Cloze test	27.34
Reading comprehension	40.34

After checking the data for unidimensionality, Bilog software was used to analyze the data. All the details on item and person statistics were obtained. The following figures show the correlations between person statistics in IRT and CTT for all three models. In other words, the Bivariate Plot provides a regression of ability on the percentage correct. A Bivariate Plot graphs the relationship between two variables that have been measured on a single sample of subjects. Such a plot permits the researcher to see at a glance the degree and pattern of relation between the two variables. On a Bivariate Plot, the abscissa (X-axis) represents the potential scores of the predictor variable and the ordinate (Y-axis) represents the potential scores of the predicted or outcome variable. Each point on the Bivariate Plot shows the X and Y scores for a single subject. This is what is meant by "Bivariate" Plot i.e. each point represents two variables.

Figure 1

Regression of ability on percentage correct for the one-parameter logistic (1PL) model



The correlation coefficient estimates the degree of closeness of the linear relationship between two variables. In this Bivariate Plot, the X-axis represents the percentage correct (CTT) and the Y-axis represents the ability (IRT). The correlation between person statistics of IRT and CTT for the one-parameter logistic (1PL) model was 0.797, which was a high correlation (figure 1).

Figure 2

Regression of ability on percentage correct for the two-parameter logistic (2PL) model

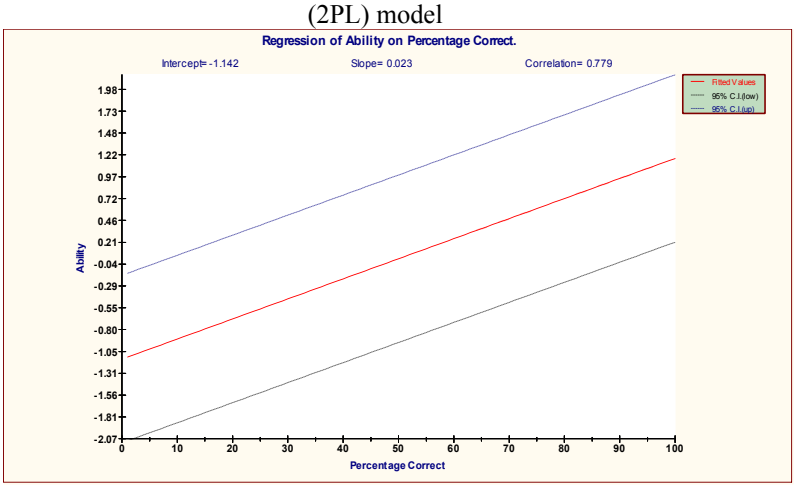
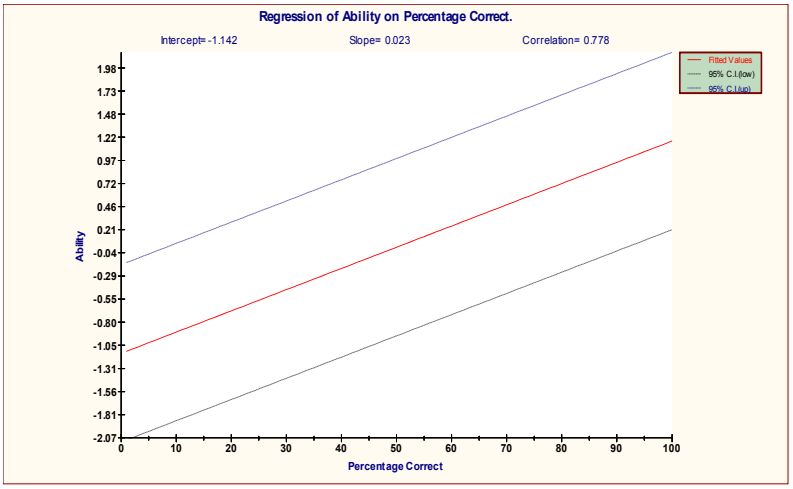


Figure 2 shows that the correlation between person statistics of IRT and CTT for the two-parameter logistic (2PL) model was 0.779, which was a rather high correlation.

Figure 3
Regression of ability on percentage correct for the three-parameter logistic 3PL model



The correlation between person statistics of IRT and CTT for the three-parameter logistic (3PL) model was 0.778, which was a rather high correlation (figure 3). Overall, there was a very high correlation between person statistics estimated through CTT and the three models of IRT. These high correlations show that CTT- and IRT-based person statistics are comparable with each other.

The following tables represent item statistics (i.e. item difficulty and item discrimination) for one of the subparts i.e. structure along with CTT and IRT models estimates for this subpart.

Table 2 provides the CTT-based item statistics for the structure subpart.

Table 2
CTT item statistics for the structure subpart

ITEM	NAME	#TRIED	#RIGHT	ITEM*TEST CORRELATION PCT	PEARSON	BISERIAL
1	ITEM1	3000.0	1577.0	52.6	0.35	0.43
2	ITEM2	3000.0	1430.0	47.7	0.29	0.37
3	ITEM3	3000.0	460.0	15.3	0.26	0.40
4	ITEM4	3000.0	1017.0	33.9	0.23	0.29
5	ITEM5	3000.0	1228.0	40.9	0.30	0.38
6	ITEM6	3000.0	2493.0	83.1	0.14	0.20
7	ITEM7	3000.0	1023.0	34.1	0.11	0.14
8	ITEM8	3000.0	1754.0	58.5	0.31	0.39
9	ITEM9	3000.0	1237.0	41.2	0.26	0.32
10	ITEM10	3000.0	2387.0	79.6	0.14	0.20

In the following three tables, which show the IRT-based item statistics, ‘slope’ is the discrimination parameter (a), ‘threshold’ is the difficulty parameter (b) and ‘asymptote’ is the guessing parameter (c). These parameters are presented for the three IRT models i.e. 1PL, 2PL and 3PL.

Table 3 provides the IRT item statistics for the structure subpart for the 1PL model. In 1PL, the value for discrimination (a) is fixed, and guessing is equal to zero; the only variable is difficulty (b).

Table 3
IRT item statistics for the structure subpart (1PL)

ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE
	S.E.	S.E.	S.E.	S.E.	S.E.
ITEM1	0.07	0.48	-0.14	0.43	0.00
	0.02*	0.01*	0.05*	0.01*	0.00*
ITEM2	-0.06	0.48	0.13	0.43	0.00
	0.02*	0.01*	0.05*	0.01*	0.00*
ITEM3	-1.13	0.48	2.37	0.43	0.00
	0.03*	0.01*	0.07*	0.01*	0.00*
ITEM4	-0.45	0.48	0.94	0.43	0.00
	0.03*	0.01*	0.05*	0.01*	0.00*
ITEM5	-0.25	0.48	0.52	0.43	0.00
	0.02*	0.01*	0.05*	0.01*	0.00*
ITEM6	1.06	0.48	-2.21	0.43	0.00
	0.03*	0.01*	0.07*	0.01*	0.00*
ITEM7	-0.44	0.48	0.93	0.43	0.00
	0.03*	0.01*	0.05*	0.01*	0.00*
ITEM8	0.23	0.48	-0.48	0.43	0.00
	0.02*	0.01*	0.05*	0.01*	0.00*
ITEM9	-0.24	0.48	0.50	0.43	0.00
	0.02*	0.01*	0.05*	0.01*	0.00*
ITEM10	0.90	0.48	-1.90	0.43	0.00
	0.03*	0.01*	0.06*	0.01*	0.00*

* STANDARD ERROR

Table 4 provides the IRT item statistics for the structure subpart for the 2PL model. In 2PL, there are two parameters or variables: discrimination (a) and difficulty (b); guessing is again equal to zero.

Table 4
IRT item statistics for the structure subpart (2PL)

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.
ITEM1	0.08	0.72	-0.11	0.58	0.00
	0.03*	0.05*	0.04*	0.04*	0.00*
ITEM2	-0.07	0.58	0.12	0.50	0.00
	0.03*	0.04*	0.04*	0.04*	0.00*
ITEM3	-1.27	0.72	1.76	0.58	0.00
	0.05*	0.06*	0.11*	0.05*	0.00*
ITEM4	-0.44	0.44	0.99	0.41	0.00
	0.03*	0.04*	0.09*	0.03*	0.00*
ITEM5	-0.26	0.59	0.44	0.51	0.00
	0.03*	0.04*	0.05*	0.04*	0.00*
ITEM6	0.98	0.29	-3.36	0.28	0.00
	0.03*	0.04*	0.42*	0.04*	0.00*
ITEM7	-0.40	0.19	2.12	0.18	0.00
	0.02*	0.03*	0.36*	0.03*	0.00*
ITEM8	0.25	0.63	-0.39	0.53	0.00
	0.03*	0.05*	0.05*	0.04*	0.00*
ITEM9	-0.24	0.49	0.49	0.44	0.00
	0.03*	0.04*	0.06*	0.04*	0.00*
ITEM10	0.84	0.28	-3.03	0.27	0.00
	0.03*	0.04*	0.39*	0.04*	0.00*

Table 5 provides the IRT item statistics for the structure subpart for the 3PL model. In the 3PL model, there are three parameters, in addition to discrimination (a) and difficulty (b); there is the guessing factor (c).

Table 5
IRT item statistics for the structure subpart (3PL)

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.
ITEM1	0.08	0.72	-0.11	0.58	0.00
	0.03*	0.05*	0.04*	0.04*	0.00*
ITEM2	-0.07	0.58	0.12	0.50	0.00
	0.03*	0.04*	0.04*	0.04*	0.00*
ITEM3	-1.27	0.72	1.76	0.58	0.00
	0.05*	0.06*	0.11*	0.05*	0.00*
ITEM4	-0.44	0.44	0.99	0.41	0.00
	0.03*	0.04*	0.09*	0.03*	0.00*
ITEM5	-0.26	0.59	0.44	0.51	0.00
	0.03*	0.04*	0.05*	0.04*	0.00*
ITEM6	0.98	0.29	-3.36	0.28	0.00
	0.03*	0.04*	0.42*	0.04*	0.00*
ITEM7	-0.40	0.19	2.12	0.18	0.00
	0.02*	0.03*	0.36*	0.03*	0.00*
ITEM8	0.25	0.63	-0.39	0.53	0.00
	0.03*	0.05*	0.05*	0.04*	0.00*
ITEM9	-0.24	0.49	0.49	0.44	0.00
	0.03*	0.04*	0.06*	0.04*	0.00*
ITEM10	0.84	0.28	-3.03	0.27	0.00
	0.03*	0.04*	0.39*	0.04*	0.00*

Table 6
Correlation for the difficulty parameter of the structure subpart

Correlations	**PCTT**	**BIRT1**	**BIRT2**	**BIRT3**
PCTT	1	-.99	-.96	-.96
BIRT1	-.99	1	.95	.95
BIRT2	-.96	.95	1	1

As can be observed in table 6, because the CTT p values were not reversed so the higher the value, the more difficult the item, the correlations between the CTT-based p values and the IRT-based item difficulty estimates were negative.

In the structure subpart, the CTT-based difficulty estimates had a high correlation with the IRT-based difficulty estimates for the three IRT models. The correlations were very high in the -.956 to -.998 range. The interesting point was that the correlation between difficulty estimates of 2PL and 3PL models was 1. Also the correlation between CTT-based p value and 1PL model difficulty (-.998) was more than the correlation between CTT-based p value and the other two models.

Table 7
Correlation for the discrimination parameter of the structure subpart

Correlations	PBISER	BISER	DIS2PL	DIS3PL
PBISER	1	.98	.94	.94
BISER	.98	1	.99	.99
DIS2PL	.94	.99	1	1
DIS3PL	.94	.99	1	1

It should be mentioned that 1PL model does not estimate item discrimination; therefore, 1PL was not included in the comparisons.

The results showed strong relationships of discrimination coefficients across measurement models in the structure subpart (table 7). There was a high correlation between CTT-based and IRT-based estimates of item discrimination, and the values were the same for the two IRT models (.939 for point-biserial and .989 for biserial) showing that there was not a difference between the 2PL and 3PL models in the estimates of item discrimination. The tables of correlations for the rest of the subparts are shown in the appendix.

Overall, concerning the correlations between the CTT-based item difficulty estimates and the IRT-based estimates, the 1PL model item

difficulty estimates provided results very similar to its CTT counterparts. For the 2PL and 3PL models, the correlations between the CTT-based item difficulty estimates and the IRT-based estimates appeared somewhat weaker, although still quite strong. It should be mentioned that when the sample is big (1000 cases or more), the estimation of item difficulty in both CTT and IRT are very close to each other and close to the population parameter, because in a way the population parameter is estimated in big samples. Consequently, with small sample sizes the difficulty parameter estimated by CTT becomes different from the population parameter.

Also considering the correlations among the difficulty estimates of the three IRT models, there was a very high correlation between 2PL and 3PL estimates. Therefore, "there seems to be little value to the IRT estimates above what CTT provides" (Courville, p. 88, 2004).

Overall, there were high correlations between the CTT-based and IRT-based 2PL and 3PL item discrimination estimates. However, in the last two subparts i.e. cloze test and reading comprehension, there were lower, albeit strong correlations between the CTT-based and the 3PL IRT-based item discrimination estimates. In other words, the item discrimination estimates from the 3PL model correlated somewhat less with CTT-based estimates than did those from the 2PL model.

With regard to item discrimination, it was also found that the correlation between biserial correlation and the two IRT models was higher than the correlation between point-biserial and the two IRT models. It is tempting to use r_{bis} rather than r_{pbi} because the former usually is larger. r_{pbi} is always less than r_{bis}, "and if the p value of the dichotomous variable is considerably different from 0.50 in either direction, r_{bis} will be much larger than r_{pbi}" (Nunnally, 1993, p. 123). So long as r_{pbi} does not equal zero, r_{bis} will be at least 25% greater than r_{pbi} computed on the same data (Millman & Greene, 1989).

r_{pbi} is the product-moment correlation between the dichotomous item scores and the criterion measure, r_{bis} is the product-moment correlation between a normally distributed latent variable underlying the right-wrong dichotomy and the criterion measure. Therefore, the difference between the measures is whether item performance is treated as a dichotomy or as a normally distributed variable. The shape of the distribution of the dichotomously scored item depends on the proportion of testees answering the item correctly; therefore, the value of r_{pbi} depends heavily on this proportion. In other words, item discrimination as measured by r_{pbi} is confounded with item difficulty, and this is what many researchers consider a major disadvantage of r_{pbi}. The discrimination of an item, as measured by r_{pbi}, changes with the ability level of the sample of testees. "Like the p value, the r_{pbi} is highly sample dependant". r_{bis} tends to be more stable from sample to sample. It is a more accurate estimate of how well the item can be expected to discriminate at some different point in the ability scale (Millman & Greene, 1989). The same point is mentioned by du Toit (2003):

> *Unlike the point-biserial, the biserial is not a product moment correlation; rather it should be thought of as a measure of association between performance on the item and performance on the test (or some other criterion). The biserial is less influenced by item difficulty and tends to be invariant from one testing situation to another –advantages the point-biserial does not possess. Also distinguishing it from its rival is the biserial correlation's assumption that a normally distributed latent variable underlies the right/wrong dichotomy imposed in scoring an item. This variable may be thought of as representing the trait that determines success or failure on the item (p. 579).*

"It is the value of r_{bis} that has the simpler, more direct relations to the ICC discrimination indicators" (Millman & Greene, 1989, p. 360). Lord and Novick (1968) mention that the extent of r_{bis} invariance is

necessarily a matter for empirical investigation, but provide some results in support of the conclusion that "biserial correlations tend to be more stable from group to group than point-biserials" (p. 340). They also show that the slope and threshold parameters of the normal ogive model for the item are functions of the biserial correlation coefficient. Therefore, item discrimination estimates of r_{bis} are closer to item discrimination estimates of the two IRT models.

Findings in the following table also justify the results to some extent. First, the chi-square values and dfs of 1PL and 2PL models for each subpart were subtracted. Then, the observed value of χ^2 was compared with the critical value in Chi-square distribution table. The same was carried out for 2PL and 3PL models.

Table 18
Chi-square values for all subparts

Subtractions ($\aleph^2$)	**$\aleph^2_{observed}$**	**Subtractions (df)**	**df**	**$\aleph^2_{critical}$**	**Probability Level**
$\aleph^2_{1PL}$-$\aleph^2_{2PL}$ Structure	190.436	df_{1PL}-df_{2PL} Structure	10	29.588	0.001*
$\aleph^2_{2PL}$-$\aleph^2_{3PL}$ Structure	0.0005	df_{2PL}-df_{3PL} Structure	10	29.588	N. Sig
$\aleph^2_{1PL}$-$\aleph^2_{2PL}$ Vocabulary	277.2774	df_{1PL}-df_{2PL} Vocabulary	20	45.315	0.001*
$\aleph^2_{2PL}$-$\aleph^2_{3PL}$ Vocabulary	0.0092	df_{2PL}-df_{3PL} Vocabulary	20	45.315	N. Sig
$\aleph^2_{1PL}$-$\aleph^2_{2PL}$ Word order	79.1954	df_{1PL}-df_{2PL} Word order	5	20.515	0.001*
$\aleph^2_{2PL}$-$\aleph^2_{3PL}$ Word order	0.0005	df_{2PL}-df_{3PL} Word order	5	20.515	N. Sig
$\aleph^2_{1PL}$-$\aleph^2_{2PL}$ Language function	60.1751	df_{1PL}-df_{2PL} Language function	5	20.515	0.001*
$\aleph^2_{2PL}$-$\aleph^2_{3PL}$ Language function	0.0021	df_{2PL}-df_{3PL} Language function	5	20.515	N. Sig
$\aleph^2_{1PL}$-$\aleph^2_{2PL}$ Cloze	272.5466	df_{1PL}-df_{2PL} Cloze	15	37.697	0.001*
$\aleph^2_{2PL}$-$\aleph^2_{3PL}$ Cloze	16.9238	df_{2PL}-df_{3PL} Cloze	15	37.697	N. Sig
$\aleph^2_{1PL}$-$\aleph^2_{2PL}$ Reading comprehension	999.0583	df_{1PL}-df_{2PL} Reading comprehension	15	37.697	0.001*
$\aleph^2_{2PL}$-$\aleph^2_{3PL}$ Reading comprehension	0.3029	df_{2PL}-df_{3PL} Reading comprehension	15	37.697	N. Sig

The null hypothesis was that there was not a significant difference between the models. First, the $\aleph^2$ value of the model with more parameters was subtracted from the $\aleph^2$ value of the model with less parameters. If $\aleph^2_{observed}$ was more than $\aleph^2_{critical}$ it showed that the model with more parameters was more suitable compared to the other one with less parameters. However, if after subtraction, $\aleph^2_{observed}$ was less than $\aleph^2_{critical}$, this showed that the two models were not significantly different

and could be used interchangeably. According to the above table, in all subparts there was a significant difference between 1PL and 2PL models at .001 level ($א^2_{observed}>א^2_{critical}$), but there was not a significant difference between 2PL and 3PL models at .001 level ($א^2_{observed}<א^2_{critical}$). Therefore, the parameters (i.e. item difficulty and item discrimination) estimated by the two models (2PL and 3PL) correlated highly with each other.

Summary

Interest in item response theory stems from two desirable features which are obtained when an item response model fits a test dataset: the item statistics are not dependent upon the particular sample of testees chosen from the population of testees for whom the test items are intended, and the expected testees' ability scores do not depend upon the particular choice of items from the total pool of test items to which the item response model has been applied. "Invariant item and examinee ability parameters, as they are called, are of immense value to measurement specialists. Neither desirable feature is obtained when the well-known and popular classical test models are used" (Hambleton, 1989. p. 4).

All three null hypotheses were accepted. The findings for this part can be summarized as follows:

1. Person statistics from CTT were comparable with those from IRT for all three IRT models.
2. Item difficulty indexes from CTT were comparable with those from all IRT models and especially from the 1PL model. Since the number of parameters in 1PL model is the least compared to the other two models (2PL and 3PL), item difficulty estimates by 1PL is closer to the CTT difficulty estimates.
3. Item discrimination indexes from CTT were somewhat less comparable with those from IRT. Although the comparability was moderately high, there was one case where the comparability was low i.e. in the cloze test subpart.

> *The lower comparability between the discrimination indexes derived from CTT and IRT implies that, in some cases, CTT and IRT may yield noticeable discrepancies*

> *with regard to which items have more discrimination power, which, in turn may lead to the selection of different items for a test, depending on which framework is used in the estimation of item discrimination (Fan, 1998, p. 375).*

Maybe this low comparability in cloze test can be justified somehow by taking into account the second assumption of IRT i.e. local independence. Since this assumption is not met completely in cloze tests, the different results maybe the consequence of not meeting this assumption.

The correlation coefficients indicated that there were considerable similarities between the item statistics obtained by CTT and IRT. Both procedures produced almost the same information regarding both item difficulties and discriminations. "However, this finding does not necessarily discredit the applicability of IRT model procedures" (Hwang, 2002, p. 18). Nunnally (1993) earlier wrote that,

> *When scores developed by ICC theory can be correlated with those obtained by the more usual approach to simply sum items scores, typically it is found that the two sets of scores correlated 0.90 or higher; thus it is really hair splitting to argue about any difference between the two approaches or any marked departure from linearity of the measurement obtained from the two approaches (p. 224).*

4. With regard to item discrimination, the correlation between biserial correlation and the two IRT models was higher than the correlation between point-biserial and the two IRT models.

5. Item difficulty and item discrimination estimates by the 2PL and 3PL models correlated very highly with each other.

Received 25 January, 2009
Accepted 25 December, 2009

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

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

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory.* New York: Holt, Rinehart and Winston.

Du Toit, M. (Ed.) (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* USA, IL: Scientific Software International.

Embreston, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Earlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/response statistics, *Educational and Psychological Measurement,* 58(3), 357-381.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* New York: Routledge.

Guilford, J. P. (1954). *Psychometric Methods.* New York: McGraw-Hill.

Hambleton, R. K. (1989). Item Response Theory: Introduction and Bibliography, *ERIC Digest No. ED 310137.*

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Cambridge, MA: Newbury House.

Hwang, D. Y. (2002). Classical Test Theory and Item Response Theory: Analytical and Empirical Comparisons, *ERIC Digest No. ED 466779.*

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Eds.), *Advances in educational research: Substantive findings, methodological developme*nt. Greenwich, CT: JAI Press.

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

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores* (with contribution by A. Birnbaum). Reading, MA: Addison-Wesley.

MacDonald, P. & 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.

Millman, J. & Greene, J. (1989). The Specification and Development of Tests of Achievement and Ability. In R. L. Linn (Eds.), *Educational Measurement.* New York: American Council on Education/Macmillan Publishing Company, 335-366.

Nunnally, J. (1993). (3rd Ed.) *Psychometric theory*. New York: McGraw-Hill.

Skaggs, G. & Lissitz, R. (1986). An exploration of the robustness of four test equating models, *Applied Psychological Measurement*, 10, 303-317.

Skaggs, G. & Lissitz, R. (1988). Effect of examinee ability on test equating invariance, *Applied Psychological Measurement*, 12, 69-82.

Stage, C. (1998a). A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT subtest WORD. (*Educational Measurement*, 29). Umea University, Department of Educational Measurement.

Stage, C. (1998b). A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT subtest ERC. (*Educational Measurement*, 30). Umea University, Department of Educational Measurement.

Stage, C. (1999). A comparison between item analysis based on item response theory and classical test theory. A study of the SweSAT subtest READ. (*Educational Measurement*, 31). Umea University, Department of Educational Measurement.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory, *Psychometrika*, 50, 399-410.

Appendix

Table 1

Correlation for the difficulty parameter for all subparts

Correlations	PCTT	BIRT1	BIRT2	BIRT3
PCTT (Structure)	1	-.99	-.96	-.96
BIRT1 (Structure)	-.99	1	.95	.95
BIRT2 (Structure)	-.96	.95	1	1
BIRT3 (Structure)	-.96	.95	1	1
PCTT (Vocabulary)	1	-.95	-.81	-.81
BIRT1 (Vocabulary)	-.95	1	.75	.75
BIRT2 (Vocabulary)	-.81	.75	1	1
BIRT3 (Vocabulary)	-.81	.75	1	1
PCTT (Word order)	1	-.99	-.97	-.97
BIRT1 (Word order)	-.99	1	.99	.99
BIRT2 (Word order)	-.97	.99	1	1
BIRT3 (Word order)	-.97	.99	1	1
PCTT (Language function)	1	-.99	-.97	-.97
BIRT1 (Language function)	-.99	1	.96	.96
BIRT2 (Language function)	-.97	.96	1	1
BIRT3 (Language function)	-.97	.96	1	1
PCTT (Cloze test)	1	-.98	-.90	-.90
BIRT1 (Cloze test)	-.981	1	.919	.894
BIRT2 (Cloze test)	-.903	.919	1	.992
BIRT3 (Cloze test)	-.899	.894	.992	1
PCTT (Reading)	1	-.999	-.960	-.960
BIRT1 (Reading)	-.999	1	.965	.965
BIRT2 (Reading)	-.960	.965	1	1
BIRT3 (Reading)	-.960	.965	1	1

Table 2
Correlation for the discrimination parameter for all subparts

Correlations	PBISER	BISER	DIS2PL	DIS3PL
PBISER (Structure)	1	.98	.94	.94
BISER (Structure)	.98	1	.99	.99
DIS2PL (Structure)	.94	.99	1	1
DIS3PL (Structure)	.94	.99	1	1
PBISER (Vocabulary)	1	.60	.65	.65
BISER (Vocabulary)	.60	1	.99	.99
DIS2PL (Vocabulary)	.65	.99	1	1
DIS3PL (Vocabulary)	.65	.99	1	1
PBISER (Word order)	1	.98	.88	.88
BISER (Word order)	.98	1	.93	.93
DIS2PL (Word order)	.88	.93	1	1
DIS3PL (Word order)	.88	.93	1	1
PBISER (Language function)	1	.97	.99	.99
BISER (Language function)	.97	1	.99	.99
DIS2PL (Language function)	.99	.99	1	1
DIS3PL (Language function)	.99	.99	1	1
PBISER (Cloze test)	1	.89	.84	.41
BISER (Cloze test)	.89	1	.98	.70
DIS2PL (Cloze test)	.84	.98	1	.74
DIS3PL (Cloze test)	.41	.70	.74	1
PBISER (Reading)	1	.98	.88	.87
BISER (Reading)	.98	1	.93	.93
DIS2PL (Reading)	.88	.93	1	1
DIS3PL (Reading)	.87	.93	1	1