The Relationship between CTT and IRT Approaches in Analyzing Item Characteristics

Nabeel Abedalaziz [1], Chin Hai Leng [2]

[1] Faculty of Education, University of Malaya, 50603Kuala Lumpur, Malaysia
nabeelabdelazeez@yahoo.com

[2] Faculty of Education, University of Malaya, 50603Kuala Lumpur, Malaysia
chin@um.edu.my

ABSTRACT

Most of the tests and inventories used by counseling psychologists have been developed using CTT; IRT derives from what is called latent trait theory. A number of important differences exist between CTT- versus IRT-based approaches to both test development and evaluation, as well as the process of scoring the response profiles of individual examinees. The purpose of this research is to compare the item difficulty and item discrimination of the Mathematical ability scale using CTT and IRT methods across 1, 2, and 3 parameters. The developed instrument was administered to tenth grade sample of N=602. The data gathered was analyzed for possible relationship of the item characteristics using CTT and IRT methods. Results indicate that the 3-parameter logistic model has the most comparable indices with CTT, furthermore, CTT and IRT models (1-parameter logistic model and 3-parameter logistic model) can be used independently or altogether to describe the nature of the items characteristics.

Keywords: *Item Difficulty, Item Discrimination, Logistic Models, Classical Test Theory, Item Response Theory*

INTRODUCTION

Classical test theory (CTT and item response theory (IRT) are widely perceived as representing two different measurement frameworks. However, few studies have empirically examined the similarities and differences in the parameters estimated using the two frameworks. Prior to exploring this issue in some detail, some readers may appreciate a brief review of related theories. Additional detail is provided elsewhere (Crocker & Algina, 1986; McKinley & Mills, 1989).

Although CTT has served the measurement community for most of this century, IRT has witnessed an exponential growth in recent decades. The major advantage of CTT is its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). Relatively weak theoretical assumptions not only characterize CTT but also its extensions (e.g., generalizability theory). Although CTT's 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. 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. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item (assuming it is dichotomously scored). This success rate of a particular pool of examinees on an item, well known as the p value of the item, is used as the index for the item difficulty (actually, it is an inverse indicator of item difficulty, with higher value indicating an easier item). The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as item discrimination, which is often 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 often computed as a point-biserial correlation coefficient.

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. For example, test equating can be accomplished empirically within the CTT framework (e.g., equipercentile equating). Similarly, empirical approaches have been proposed to accomplish item-invariant measurement (e.g., Thurstone absolute scaling) (Englehard, 1990). It is fair to say that, to a great extent, although there are some issues that may not have been addressed theoretically within the CTT framework, many have been addressed through ad hoc empirical procedures. IRT, on the other hand, is more theory grounded and models the probabilistic distribution of examinees' success at the item level. As its name indicates, IRT primarily focuses on the item-level information in contrast to the CTT's primary focus on test-level information. The IRT framework encompasses a group of models, and the applicability of each model in a particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items. For test items that are dichotomously scored, there are three IRT models, known as three-, two-, and one-parameter IRT models.

Although tests have always been composed of multiple items, item response theory (IRT) takes a much more item-level focus than classical test theory (CTT), which tends to focus more on test-level indices of performance (e.g., the overall reliability coefficient, or standard error, of a scale). In particular, the focus on estimating an ICC for each item provides an integrative, holistic view of the performance of each item that is not readily available when using CTT-based methods to develop or examine a test. That is, although CTT can quantify the total-sample difficulty (e.g., as a p value) or discrimination (e.g., as an item-total biserial correlation) for an item, it lacks an effective means for simultaneously combining and presenting this information (including the role of guessing, or other factors that might lead to a nonzero lower asymptote) in an easily-used format.

With respect to test scoring, IRT-based tests – especially those based on the 2- or 3- parameter models – offer considerable advantages over the "number right" scoring methods typically used in CTT-based tests. Specifically, when estimating an examinee's score using IRT, we can simultaneously consider the following sources of information: (a) which items were answered correctly/incorrectly (or in the keyed vs. non-keyed direction); and (b) for each of those items, the difficulty, discrimination, and nonzero lower-asymptote parameters of the item. This offers the potential to produce better estimates of the θ scores, to produce quantitative estimates of the "quality" or likelihood of any given observed response profile (termed appropriateness indices; e.g., Drasgow, 1987), and to assess the degree to which the given IRT model provides a good "fit" to the pattern of responses produced by the individual in question..

Classical Test Theory (CTT) test score state that an examinee's observed score consists of his/her true score plus error. IRT has a similar interest in determining an examinee's true score (latent trait score). However, CTT approaches are limited in that examinee ability is defined in terms of a particular test, and the difficulty of that test is determined by the ability of the examinees who take it. This circularity of item and examinee characteristics in CTT branches into the estimation of reliability and validity as well because the test and item characteristics change as the examinee pool changes. Item Response Theory models, contrary to CTT models, are falsifiable in that they may or may not be appropriate for a particular data set (Hambleton et al., 1991). IRT models do not suffer from the limitations of CTT because item and ability parameters are invariant under a linear transformation (i.e., it is possible to change the means and variance estimates for different subgroups so that they lie on the same metric). Estimates of item parameters obtained from different examinee groups will be the same, and estimates of examinee ability do not depend on the pool of items administered (except for sampling or measurement errors; Hambleton et al., 1991).

Lastly, Classical Test Theory is limited in that it can only provide test level information. There is no consideration of how examinees perform on individual items (other than via statistics such as the item p value). It is sometimes essential to be able to design tests with items targeted toward specific ability levels. IRT models allow a test developer to design items that, for example, discriminate well among high ability examinees (Hambleton et al., 1991). In short, IRT models, because they provide item level information, are far superior to CTT models for many testing applications, especially those that seek to examine the performance of individual test items.

IRT ASSUMPTIONS

There are two primary postulates of IRT: (a) Examinee performance on a test item is a function of latent traits, or abilities; and (b) the graphical relation between examinees' latent traits and their probabilities of answering an item correctly is in the form of a monotonically increasing function called an item characteristic curve (ICC). In other words, item performance depends on latent traits (e.g., ability), and as the level of the latent trait increases, the probability of a correct response either increases or stays the same (Hambleton et al., 1991). In IRT models, the

underlying latent trait is referred to as theta (q), which is conceptually similar to a "true score" in Classical Test Theory.

The graph of an item characteristic curve has, on its x-axis, q (expressed typically as a Z-score ranging from -3 to +3), and on its y-axis, the probability of a correct response (PCR). There are also several assumptions about the data to which IRT models are applied. The first assumption is that of unidimensionality, that one ability (latent trait) is measured by a test. In order for this assumption to be adequately met in an IRT model, a set of test data must consist of a "dominant" factor from which overall test performance results (Hambleton et al., 1991). Local independence, while related to unidimensionality, is the assumption that when ability (the latent trait) is held constant, there should be no relation between examinees' responses to different items (Hambleton et al., 1991). In other words, the underlying latent trait the test purports to measure should be the only factor that has an overall influence on responses to test items, and when that latent trait is statistically controlled, there should be nothing consistently affecting item performance, and thus, the items should be uncorrelated (independent).

When there is an adequate fit between an IRT model and a set of test data, there are several desirable results such as test-free measurement. Test-free measurement implies that the estimates of examinee ability are assumed to be the same even if a different set of items is used (barring measurement errors), and item parameter estimates will be identical for different groups of examinees (except for sampling errors; Hambleton et al., 1991). This property of invariance of item and ability parameters is one of the advantages of IRT models.

IRT MODELS

The one-parameter logistic model, explains the relationship between levels of the ability and probability of a correct response on the item in terms of the difficulty of the item. An item's b parameter (difficulty) is the point on the ability scale corresponding to the location on the ICC where the probability of a correct response is 0.5 (Hambleton et al, 1991).

The two –parameter logistic model makes use of the b parameter (item difficulty) just as in the one-parameter model, but adds an additional element which indicates how wills an item separates examinees into different ability levels. The a parameter used in the two-parameter model is called the item discrimination parameter and is equal to the slope of the ICC when it is at its steepest (Hambleton et al , 1991).

The three-parameter logistic model builds upon the two-parameter model by adding pseudo-chance-level parameter c. The c parameter is the value of the lower asymptote of the item characteristic curve and is indicative of the probability that an examinee with a very low ability score would answer an item correctly.

Studies linking CTT and IRT item characteristics have been done and have shown signs of positive indications of a relationship that exists (Hernandez, 2009; Adedoyin, Nenty, and Chilisa, 2008; Nukhet, 2002; Fan, 1998). However, local literature has yet to replicate the studies and results.

The main objectives of the present study was to analyze the item characteristics of a Mathematical ability Scale using both CTT and IRT methods and to check if both methods are comparable and can used independently or interchangeably. Specifically, the present study sought answers to the following questions: (1) How comparable are the CTT-based and IRT-based item *difficulty* estimates? (2) How comparable are the CTT-based and IRT-based item *discrimination* estimates?

METHOD

Participants

The scale of the mathematical ability was applied during the last quarter of the school – year 2009/2010 to samples of 602 students- males and females- from the tenth grades.

Instrument

A mathematical ability scale was developed in order to measure two components of the mathematical ability: numerical ability, and spatial ability. The primary form of the scale was tried out to a sample of 144 students-males and females, chosen from the eighth, ninth and tenth grades to make sure that the items of the scale are clear and are understood by those who were tested, and to recognize the levels of difficulty, discrimination, and the effectiveness of the detractors of the items. Accordingly, the final version of the scale consisted of 74 items (37 items for numerical ability subtest, and 37 items for spatial ability).

The scale of the mathematical ability was applied during the last quarter of the school – year 2009/2010 to sample of 602 students- males and females- from the eighth, ninth and tenth grades. The item analysis revealed levels of difficulty from 0.27 to 0.92 and levels of discriminate ability from o.30 to o.56. Besides, it revealed that the detractors were reversal to the item discriminate.

Data about validity of the scale were collected through six methods: Internal consistency, item analysis, Logical judgment, Factor analysis, analysis of variance (One way MANOVA) and correlation with mathematical achievement.

In order to collect data about the reliability of the scale, the following four approaches were used: Cronbach alpha method, parallel forms reliability and test – retest method. Confirmatory Factor Analysis reveals that the test measures a single trait (unidimensionality).

DATA ANALYSIS

Data gathered were then analyzed using SPSS version 16, BILOG-MG, and Microsoft Excel version 2007. Classical Test Theory analysis was done using the SPSS program. The software automatically generated the following: item difficulty, item discrimination, and point biserial correlation (r_{pb}). To prepare the data for correlation with the IRT parameters, difficulty and (r_{pb}) had to be transformed into a Z (normal) distribution, Δ and Z respectively (Fan, 1998; Anastasi, 1988; Holland and Thayer, 1985).

IRT parameters were obtained using the BILOG-MG program. The program generated the item difficulty (b - parameter) and item discrimination (a - parameter) for one, two, and three parameter logistic model. Goodness of Fit tests were used to examine how many items had an ICC that fitted the three models. 26 items did not fit the 1PL model, 18items did not fit the 2PL model and 3 items did not fit the 3PL model.

Pearson product moment correlation was then used to determine the relationship between the variables being studied, and the significance of Pearson product moment correlation coefficients were tested. CTT difficulty was correlated with the b parameters of IRT (1-parameter, 2-parameter, and 3-parameter) logistic models. CTT point biserial correlation was correlated with the parameters of IRT (2-parameter, and 3-parameter) logistic models. The coefficient of determination (R^2) was obtained.

RESULTS

Table 1 shows the mean and standard deviation values of the Numerical ability and spatial ability subtests when classified into CTT and IRT. Comparison of CTT Difficulty and Discrimination scores show that the item difficulty index of both subtests are of average difficulty with the Spatial ability subtest slightly higher or easy than the numerical ability subtest. The CTT item discrimination values for both subtests (spatial, and numerical) indicates their reasonable discrimination between high and lows corers. The spatial ability subtest also shows better discrimination compared to the numerical ability subtest for the 3-parameter logistic model, the spatial ability subtest shows lower discrimination compared to the numerical ability subtest for the 2-parameter logistic model. The IRT Difficulty parameters for the 1-parameter logistic model generally have the lowest values for both spatial and numerical subtests. This indicates that the 1-parameter logistic model provides the lowest possible item difficulty index. Conversely, the 3-parameter logistic model has the highest values. On the other hand, item discrimination as measured in IRT reveal that the 2-parameter logistic model provides the lowest parameter values.

Table 1. Mean and Standard Deviation of item difficulty and item discrimination for subtests

		Numerical ability				Spatial ability			
		CTT	**IRT**			**CTT**			**IRT**
			1-P	2-P	3-P		1-P	2-P	3-P
Item difficulty	**Mean**	0.58	-0.48	-0.30	0.15	0.54	-0.37	0.06	0.52
	S.D	0.15	0.83	0.41	0.71	0.18	1.15	1.62	1.10
Item discrimination	**Mean**	0.47	-	1.24	2.29	0.46	-	0.57	1.65
	S.D	0.05	-	0.44	1.29	0.03	-	0.21	0.56

Note. S.D: Standard deviation, CTT: Classical test theory, IRT: item response theory, 1-P: One parameter logistic model, 2-P: Tow parameter logistic model, 3-P:Three parameter logistic model.

Table 2 reveals that generally, there is a significant positive and linear relation that exists between CTT and IRT in terms of item difficulty and item discrimination. However, there is no significant relation between items discrimination measured by CTT and the discrimination parameters as measured by the 3-parameter model for the spatial ability subtest. The correlation coefficients between CTT and IRT models in term of item difficulty are from 0.60 to 0.95, and the correlation coefficients between CTT and IRT models in term of item discrimination are from 0.85 to 0.93. Table 2 shows that there exists a variation in the coefficient of determination values of the three IRT models when correlated versus the CTT item difficulty and item discrimination. The 3-parameter model has the largest (R^2) value of 0..90, and the 1-parameter model has the lowest (R^2) value of 0.36 in term of item difficulty for the numerical subtest, and the 3-parameter model has the largest (R^2) value of 0.88 and the 1-parameter model has the lowest (R^2) value of 0.37 in term of item difficulty for the spatial subtest. Also, the 3-parameter model has the largest (R^2) value of 0.86, and the 2-parameter model has the lowest (R^2) value of 0.62 in term of item discrimination for the numerical subtest, and The 3-parameter model has the largest (R^2) value of 0.86 and the 2-parameter model has the lowest (R^2) value of 0.60 in term of item discrimination for the spatial subtest.

Table 2. Correlations of Difficulty and Discrimination on Logistic Parameters (N=600)

	Numerical ability			Spatial ability		
	1-P	**2-P**	**3-P**	**1-P**	**2-P**	**3-P**
Item difficulty	0.60**	0.81**	0.95**	0.61**	0.82**	0.94**
Item discrimination	-	0.80**	0. 93**	-	0.80*	0.93**

**Significant at α=0.01

DISCUSSION

In the theory of measurement, there are two competing measurement frameworks, classical test theory and item response theory. The present study empirically examined how the item statistics behaved under the two competing measurement frameworks**.** Regarding to the results, it is evident that there is a significant relationship between the CTT and IRT approaches in analyzing the item characteristics of the mathematical ability scale (positive linear relations). Results further revealed that when items are categorized from easy to hard item difficulty (in CTT), it would also correspond to almost the same IRT classification of item difficulty. The same can be said for item discrimination categorization between CTT and IRT approaches.

Results revealed that the one- parameter logistic model shows lower significant relationship to CTT in term of item difficulty, and the three-parameter logistic model shows higher significant relationship to CTT in term of item discrimination and item difficulty, whereas, the two-parameter logistic model reveals lower significant relationship to CTT in term of item discrimination. As such, the three-parameter logistic model has the most comparable indices with CTT. These findings seem to be consistent with previous researches (e.g. Fan, 1998; Nukhet, 2002). Nukhet (2002) reported that the three-parameter logistic model has the most comparable indices with CTT. Fan (1998) indicated that all the three IRT models are comparable with CTT. In the contrast, some researchers found that the two-parameter logistic model has the most comparable indices with CTT (Hernandeze, 2009; Adedoyin, Nenty, & Chilisa, 2008; Nukhet, 2002; Fan, 1998). Hernandeze (2009) established that CTT and IRT can be used independently or altogether to describe the nature of the items.

In the present study, CTT approaches and IRT logistic models can be used independently or altogether to describe the item characteristics. These results corroborate results reported by Lawson (1991), Fan (1998), Stage (1999), and MacDonald and Paunonen (2002) all indicating that CTT and IRT measurement theories often produce quite similar results.

CONCLUSION

The results of this study are supports Nunnally's (1979) assertion that "when scores developed by item response 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 .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). Overall, the results of this study indicate that CTT-based and IRT-based estimates, for the three models, are quite similar. The CTT-based item difficulty estimates and the three models IRT item difficulty

estimate provided very similar results.

The item discrimination should be concluded to be unequal among the items which lead to the conclusion that we should have that parameter in our model. That is, the one-parameter logistic model is less suitable than the two-parameter logistic model or the three-parameter logistic model .Also, guessing parameter should be in the model since some examinees with low ability tend to guess the correct answer on the most difficult items. This result suggests that the three-parameter logistic model model is preferable over the one-parameter logistic model and the two-parameter logistic model.

Goodness of Fit tests should be used to examine how many items had an item characteristic curve that fitted the three models. In the present study, 26 items did not fit the 1PL model, 18 items did not fit the 2PL model and 3 items did not fit the 3PL model. This result suggests that the 3PL model is preferable to the other models. The 2PL model and the 3PL model are, in general, to be preferred over the 1PL model. The results suggest that guessing should be part of the model. The overall conclusion is that the 3PL model is most suitable to model the items and higher correlate with CTT statistics. Finally, if both CTT and IRT are used when evaluating items, different dimensions of information are obtained since both CTT and IRT add valuable information about the test.

REFERENCES

Adedoyin, O. O., Nenty, H.J, & Chilisa, B. (2008). *Investigating The Invariance of Item Difficulty Parameter Estimates Based on CTT and IRT*. Educational Research and Review, 3 (2), 83-93.

Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York: Macmillan.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart & Winston.

Drasgow, F. (1987). *Study of The Measurement Bias of Two Standardized Tests*.

Englehard, G., Jr. (1990, April). *Thorndike, Thurstone and Rasch: A Comparison of Their Approaches to Item-invariant Measurement*. Paper presented at the annual meeting of the American Educational Research Association, Boston. (ERIC Document Reproduction Services No. ED 320 921).

Fan, X. (1998). *Item Response Theory and Classical Test Theory: An Empirical Comparison of Their Item/Person Statistics*. Educational and Psychological Measurement, 58, 357-385.

Hambleton, R. K., & Jones, R. W. (1993). *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development*. Educational Measurement: Issues and Practice, 12(3), 3847.

Hambleton, R. K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hernandez, R. (2009). *Comparison of The Item Discrimination and Item Difficulty of the Quick-Mental Aptitude Test using CTT and IRT Methods*. The International Journal of Educational and Psychological Assessment, Vol. 1, Issue 1, pp. 12-18.

Holland, P. W. & Thayer, D. T. (1985). *An Alternative Definition of The ETS Delta Scale of Item Difficulty*. Educational testing Service, Technical report (85-64)/ Research Report (85-43). Journal of Applied Psychology, 72, 19-29.

Lawson, S. (1991). One Parameter Latent Trait Measurement: Do The Results Justify the Effort?. In B. Thompson (Ed.), *Advances in Educational Research: Substantive Findings, Methodological Developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

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.

McKinley, R., & Mills, C. (1989). *Item Response Theory: Advances in Achievement and Attitude Measurement*. In B.Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 71-135). Greenwich, CT: JAI.

Nukhet, C. (2002) *A Study of Raven Standard Progressive Matrices Test's Item Measures Under Classic and Item Response Models: An Empirical Comparison*. Ankara University, Journal of Faculty of Educational Science, 35 (1-2), 71-79.

Nunnally, J. (1978). *Psychometric Theory* (2nd ed.). New York: McGraw-Hill.

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 No 31). Umea University, Department of Educational Measurement.