Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2014/04/jls.htm
2014 Vol. 4 (S4), pp. 424-435/Zoghi and Valipour
Research Article

A COMPARATIVE STUDY OF CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY IN ESTIMATING TEST ITEM PARAMETERS IN A LINGUISTICS TEST

Masoud Zoghi and *Valeh Valipour
Department of English Language Teaching, College of Humanity, Ahar Branch, Islamic Azad University, Ahar, Iran
**Author for Correspondence*

ABSTRACT
This study was an attempt to assess the comparability test items parameter estimates between the classical test theory (CTT) and the item response theory (IRT) models. To that end, the researchers selected 40 participants from Islamic Azad University of Tonekabon, Iran. The participants took two linguistics tests. To analyze the collected data, SPSS and IRTPRO computer software programs were used. To estimate the test items parameters in terms of item difficulty, item discrimination, and the responses given by the students to each item, CTT and IRT (2PL) models were used. Results suggested that CTT and IRT test items parameters are comparable.

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

INTRODUCTION
Mainly, the focus of test development is on the quality of test items and how examinees respond to them. Considering psychometric theories, there are two prevailing approaches to determining the validity and reliability of a test – Classical Test Theory (CTT) and Item Response Theory (IRT). These approaches have some similarities and differences that are briefly touched upon below. As an initial step, we posited our discussions on a brief definition of each approach.

Classical Test Theory
CTTis a psychometric theory for predicting the results of psychological testing and is concerned with the difficulty of items and the ability of test-takers. The purpose of classical test theory is to clarify as well as to improve the reliability of psychological tests. CTT is considered as a model related to true score theory. According to Margno (2009),"CTT is based on the true score model, which depends on examinee's aggregate score in a test and therefore does not permit a consideration of examinees' responses to any specific item, providing no basis to predict how a given examinee will perform on a particular test item."
Novick (1966) describes CTT as:"A theory assumes that each person has a *true score* that would be obtained if there were no errors in measurement. A person's true score is defined as the expected number-correct score over an infinite number of independent administrations of the test. Unfortunately, test users never observe a person's true score, only an *observed score*, *X*. It is assumed that *observed score = true score* plus some *error*: Classical Test Theory is concerned with the relations between the three variables X, T, and E in the population."
According to Hambleton and Jones (2005), the advantages obtainable through the application of classical test models to measurement problems are: (a) it can be performed with smaller samples of test-takers as representatives of a population, (b) it employs relatively simple mathematical procedures and conceptually straightforward model parameter estimations, and (c) it is considered as a weak model asits assumptions are easily met by traditional testing procedures.
Although CTT has been proved to be effective in test development, some limitations have still been associated with it. Item difficulty and item discrimination as the main aspects used in this theory is sample dependent, i.e. the obtained results are so much dependent on samples for being interpreted. CTT is also test dependent or test based. The difficulty of tests affects the test scores gained and the true score model CTT on which it is based, and it doesn't let any place for examinees' responses to any particular item. Therefore, it can be no prediction about how an examinee performs on a special test item. In

addition to the above mentioned disadvantages, in CTT, there is no separation between test-takers' characteristics and test characteristics, and they have to be interpreted in the context of each other. Also, the definition of reliability is problematic as it is defined "the correlation between test scores on parallel forms of a test - the problem is that there are differing opinions of what parallel tests are. Various reliability coefficients provide either lower bound estimates of reliability or reliability estimates with unknown biases" (Hambleton *et al.,* 1991). Another disadvantage in CTT is the standard error of measurement which is considered the same for all test-takers while according to Hambleton (1991), each test taker's score on any test shows unequal difference of his /her ability. The last disadvantage of CTT is being test oriented rather than item oriented, so CTT cannot be useful in predicting a test-takers performance on a test.

Item Response Theory (IRT)

Item response theory consists of any model "relating the probability of an examinee's response to a test item to an underlying ability" (HMIRT, p. v). It is called latent trait theory attempting to predict observations from places on latent variable.

According to (Reeve, 2002) "IRT item parameters are not dependent on the sample used to generate the parameters, and are assumed to be invariant (within a linear transformation) across divergent groups within a research population and across populations."

Hambleton (1991) points out the two basic postulates of IRT:

a. The performance of an examinee on a test item can be predicted or explained by a set of factors called traits, latent traits, or abilities,

b. The relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases.

A Comparison of Classical Test Theory and Item Response Theory

IRT makes assumptions, findings and characteristics of errors more strongly in comparison with CTT as well as its model-based nature; it also has many advantages over analogous CTT findings which lead to important practical results. CTT is simple to compute and elaborate while scores according to IRT need complex computation procedures but its advantage is comparing meaningfully the difficulty of an item and the ability of a person on a test. Also the parameters of IRT models are not samples or test dependent, therefore, IRT can provide more flexibility in situations with different samples or test forms, and its findings are basic for computerized adaptive testing.

Purpose of the Study

The purpose of the present study is to assess the comparability of test items parameter estimates between classical test theory (CTT) and Item response theory (IRT) models as well as a significant relationship between the IRT and CTT test items estimation, in estimating test item parameters in linguistics test items in Islamic Azad University, Tonekabon, Iran. This study was guided by the following research questions:

1. What are the item parameter measurements of participants' responses to linguistics test items based on CTT model?
2. What are the item parameter measurements of participants' responses to linguistics test items based on IRT (2PL) Rasch model?
3. Is there any particular statistical comparability between the CTT and IRT item parameter measurements of participants' responses to linguistics test items?

MATERIALS AND METHODS

Methodology

Participants of the Study

The participants of the present study consisted of 40 university-level seniors majoring in English Language Translation in a university located in the province of Mazandaran – Iran. Among the student population (N=112) in the selected university who expressed their acceptance to participate in the current

study, the sample (N= 40) was randomly selected, both male and female in the age range of 21-25. All the students had equal chances to be selected.

Instrument

To collect the desired data and information, two linguistics tests were developed. Each of these tests consisted of 20 multiple choice questions. The given time for responding the items was 60 minutes.

Data Analysis

To analyze IRT, the model fit and uni-dimensionality goodness of fit tests were used to examine how many items fitted the 2PL models and an exploratory factor analysis, then confirmatory factor analysis were performed to test. They were then subjected to IRT (Rasch) model to find the item parameter estimates of the test items in terms of item difficulty and item discrimination. The responses of the 40 students selected from among 112 ones were analyzed to find CTT and IRT test items parameter estimates.

For CTT item parameter estimates, SPSS program version 16 was used to estimate the parameter estimates for CTT item difficulty, item discrimination, transforming the item parameters of both CTT and IRT into *z*-scores and point biserial correlation (*rpb*) between CTT and IRT.

For IRT item parameter estimates, IRTPRO 2.1 software was used to estimate the item difficulty and item discrimination parameters using the IRT (Rasch model).

First of all by performing descriptive statistics, mean, standard deviation of each items were estimated (Table 1), then to estimate factor analysis and showing factorability exploratory and confirmatory factor analysis were done (Table 2, 3). Pearson's coefficient of correlation between each extracted item and the total test score revealed Correlation was significant at $p < 0.01$, 2-tailed as well as at $p < 0.05$, 2-tailed (Table 4). The results of factor analysis was shown in table5, communalities_ It may be seen from Table 5, out of 20 items, only 8 items Viz. 1, 2, 5, 7, 8, 9, 11, and 13 have shown a significant correlation with total test score and Table 6, Total Variance Explained by the result of factor analysis through Extraction Method (Principal Component Analysis). Then Scree Plot for the Eigen values confirms the results gained from the mentioned table (Figure 1).

Table 1: Descriptive Statistics

	Mean	Std. Deviation	Analysis N
q1	.7250	.45220	40
q2	.6250	.49029	40
q3	.4250	.50064	40
q4	.5250	.50574	40
q5	.3500	.48305	40
q6	.2000	.40510	40
q7	.4000	.49614	40
q8	.4250	.50064	40
q9	.3500	.48305	40
q10	.7750	.42290	40
q11	.5750	.50064	40
q12	.3750	.49029	40
q13	.5750	.50064	40
q14	.4750	.50574	40
q15	.5000	.50637	40
q16	.5000	.50637	40
q17	.3250	.47434	40
q18	.2500	.43853	40
q19	.6500	.48305	40
q20	.6250	.49029	40

Table 2: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**0.365**
Bartlett's Test of Sphericity	Approx. Chi-Square	223.600
	df	190
	Sig.	0.048

Table 3: The number of items fitting each factor

Factors	**Items fitting the model**
Factor1	1,2,4,7
Factor2	6,10,12,18

Table 4: Pearson's coefficient of correlation between each extracted item and the total test score

		q1	**q2**	**q4**	**q6**	**q7**	**q10**	**q12**	**q18**	**Total Scores**
*q*1	Pearson Correlation	1	.448**	.311	.308	.389*	-.064	-.217	.097	.478**
	Sig. (2-tailed)		.004	.051	.053	.013	.696	.179	.552	.002
	N	40	40	40	40	40	40	40	40	40
*q*2	Pearson Correlation	.448**	1	.297	.129	.632**	-.046	-.253	.089	.433**
	Sig. (2-tailed)	.004		.062	.427	.000	.776	.115	.583	.005
	N	40	40	40	40	40	40	40	40	40
q4	Pearson Correlation	.311	.297	1	.100	.061	-.513**	-.090	-.029	.288
	Sig. (2-tailed)	.051	.062		.539	.707	.001	.579	.859	.071
	N	40	40	40	40	40	40	40	40	40
q6	Pearson Correlation	.308	.129	.100	1	-.026	.269	-.387*	-.289	.148
	Sig. (2-tailed)	.053	.427	.539		.876	.093	.014	.071	.362

	N	40	40	40	40	40	40	40	40	40
q7	Pearson Correlation	.389*	.632**	.061	-.026	1	-.049	.000	.000	.367*
	Sig. (2-tailed)	.013	.000	.707	.876		.765	1.000	1.000	.020
	N	40	40	40	40	40	40	40	40	40
q10	Pearson Correlation	-.064	-.046	-.513**	.269	-.049	1	-.201	-.104	-.095
	Sig. (2-tailed)	.696	.776	.001	.093	.765		.214	.524	.558
	N	40	40	40	40	40	40	40		
q12	Pearson Correlation	-.217	-.253	-.090	-.387*	.000	-.201	1	.149	.179
	Sig. (2-tailed)	.179	.115	.579	.014	1.000	.214		.359	.268
	N	40	40	40	40	40	40	40	40	40
q18	Pearson Correlation	.097	.089	-.029	-.289	.000	-.104	.149	1	.118
	Sig. (2-tailed)	.552	.583	.859	.071	1.000	.524	.359		.469
	N	40	40	40	40	40	40	40	40	40
TotalScores	Pearson Correlation	.478**	.433**	.288	.148	.367*	-.095	.179	.118	1
	Sig. (2-tailed)	.002	.005	.071	.362	.020	.558	.268	.469	
	N	40	40	40	40	40	40	40	40	40

***. Correlation is significant at the 0.01 level (2-tailed).*
***. Correlation is significant at the 0.05 level (2-tailed).*

*According to Table 1, since the magnitude of KMO (0.365) is less than *0.5*,

it indicates that the power of data for factorability is weak, but, since the amount of Bartlett's Test of Sphericity =23.60 is significant at $p < 0.05$ level, the power of data for factorability is confirmed.

Table 5: Communalities

	Initial	Extraction
q1	1.000	.697
q2	1.000	.837
q3	1.000	.738
q4	1.000	.844
q5	1.000	.748
q6	1.000	.723
q7	1.000	.815
q8	1.000	.723
q9	1.000	.429
q10	1.000	.815
q11	1.000	.839
q12	1.000	.713
q13	1.000	.715
q14	1.000	.671
q15	1.000	.764
q16	1.000	.599
q17	1.000	.746
q18	1.000	.647
q19	1.000	.626
q20	1.000	.628

Extraction Method: PrincipalComponent Analysis
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Table 6: Total Variance Explained by the result of factor analysis

	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	2.734	13.670	13.670	2.734	13.670	13.670	2.237	11.187	11.187
2	2.185	10.926	24.596	2.185	10.926	24.596	1.953	9.767	20.954
3	1.933	9.664	34.259	1.933	9.664	34.259	1.935	9.673	30.627
4	1.893	9.466	43.725	1.893	9.466	43.725	1.846	9.230	39.857
5	1.703	8.513	52.238	1.703	8.513	52.238	1.645	8.223	48.079
6	1.417	7.087	59.325	1.417	7.087	59.325	1.588	7.941	56.021
7	1.339	6.697	66.022	1.339	6.697	66.022	1.563	7.816	63.837
8	1.113	5.564	71.586	1.113	5.564	71.586	1.550	7.750	71.586
9	.986	4.932	76.519						
10	.879	4.393	80.911						
11	.729	3.646	84.557						
12	.677	3.384	87.941						
13	.556	2.782	90.723						
14	.463	2.316	93.038						
15	.392	1.960	94.999						
16	.293	1.464	96.462						
17	.224	1.119	97.581						
18	.199	.994	98.575						
19	.183	.914	99.489						
20	.102	.511	100.000						

Extraction Method: Principal Component Analysis.

Results revealed eight extracted components that have high factorial load. The total variance explained by theses components have shown in Table 3. The two first factors have the greatest factorial load. The two first Eigen values were 2.734 and 2.185 greater than the next six Eigen values (1.933, 1.893, 1.703, 1.417, 1.339 and 1.113). The first factor explained 13.67% of the variance in the data set. The second factor explained 10.926% of the remaining variance. The rest of the variance was explained by the other six factors each having a percentage of variance between 5.5 and 9.7.

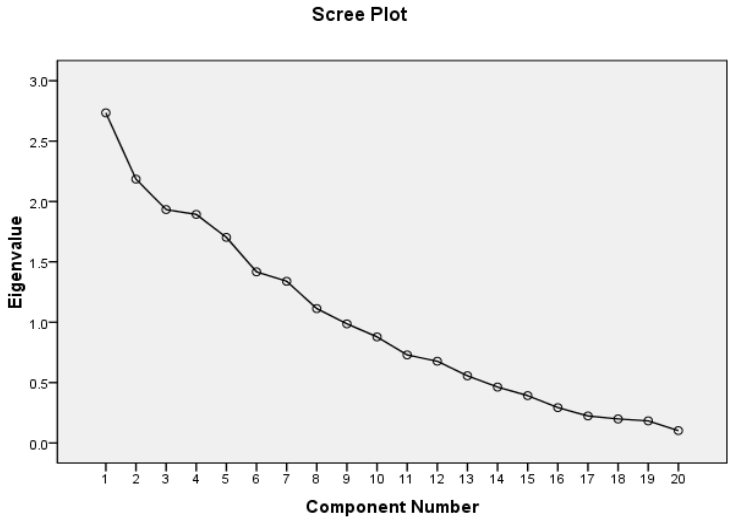

Figure 1: Scree Plot for the Eigen values

The Scree plot confirms the information presented in Table 6. It may be seen from the above plot the Eigen values of the extracted components are more than 1 which are suitable.

Table 7: CTT Model Item Parameter Estimates

ITEMS	CTT *p*-values	CTT *a*-values
1	.27	.45
2	.375	.35
4	.475	.25
6	.8	.00
7	.6	.3
10	.225	.00
12	.625	.1
18	.75	.3

IRT Rasch Model Estimates

Table 8: 2PL Model Item Parameter Estimates, logit: aθ + c or a (θ – b)

Item	Label	*a*		*s.e.*	*c*		*s.e.*	*b*	*s.e.*
1	item1	2	0.66	0.50	1	0.56	0.36	-0.85	0.77
2	item2	4	0.34	0.39	3	0.21	0.33	-0.61	1.17
3	item3	6	-0.26	0.40	5	0.63	0.34	2.41	3.80
4	item4	8	0.60	0.48	7	0.11	0.34	-0.18	0.58
5	item5	10	0.08	0.39	9	0.10	0.32	-1.20	6.70
6	item6	12	0.32	0.40	11	0.52	0.33	-1.64	2.21
7	item7	14	0.21	0.39	13	-0.10	0.32	0.49	1.80
8	item8	16	-6.77	2.79	15	5.51	2.59	0.81	0.21
9	item9	18	0.47	0.43	17	-0.00	0.33	0.00	0.70
10	item10	20	2.00	1.02	19	0.33	0.51	-0.16	0.25
11	item11	22	1.58	0.84	21	0.74	0.51	-0.47	0.32
12	item12	24	0.53	0.43	23	0.21	0.34	-0.40	0.70
13	item13	26	-0.65	0.55	25	1.20	0.42	1.84	1.46
14	item14	28	-0.84	0.49	27	-0.47	0.38	-0.56	0.51
15	item15	30	-0.34	0.40	29	0.53	0.34	1.53	1.93
16	item16	32	-0.36	0.43	31	0.00	0.32	0.00	0.90
17	item17	34	-0.33	0.39	33	0.00	0.32	0.00	0.97
18	item18	36	-0.44	0.41	35	-0.42	0.34	-0.97	1.13
19	item19	38	-0.72	0.49	37	0.82	0.39	1.13	0.80
20	item20	40	-0.05	0.38	39	0.30	0.32	6.11	47.75

Table 9: Summed-Score Based Item Diagnostic Tables and X^2s S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	Probability
1	item1	4.85	4	0.3047
2	item2	8.89	4	0.0637
3	item3	5.04	4	0.2844
4	item4	2.90	3	0.4080
5	item5	4.73	3	0.1937
6	item6	5.57	4	0.2351
7	item7	11.72	3	0.0084
8	item8	2.89	2	0.2369
9	item9	4.10	3	0.2517
10	item10	5.79	3	0.1218
11	item11	6.09	4	0.1922
12	item12	5.07	3	0.1662
13	item13	3.77	2	0.1529
14	item14	8.36	4	0.0791
15	item15	4.00	3	0.2631
16	item16	4.62	4	0.3302
17	item17	6.92	3	0.0744
18	item18	4.39	3	0.2235
19	item19	7.80	4	0.0990
20	item20	10.93	4	0.0273

**The chi-square goodness of fit analysis showed that 19 items fitted the 2PL model*

Table 10: Result of the dependent t-test between CTTzp and IRTzb ;CTTza and IRTza,

	Paired Differences							
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	*t*	*df*	Sig. (2-tailed)
CTTzp and IRTzb	-3.00000	6.00000	1.00000	-6.00000	.00000	-2.000	19	.024
CTTza and IRTza	3.00000	6.00000	1.00000	.00000	6.00000	2.000	19	.024

RESULTS AND DISCUSSION

Results

This study was done to find out the comparability between the estimation of test items parameters in terms of Classical test theory (CTT) and Item response theory (IRT) Rasch models 2PLM, using two different linguistics tests, each consisted of 20 items done by the same participants, 40 university students.

The methods used to estimate and explore uni-dimensionality in this study were first exploratory factor analysis to identify the underlying relationships between measured variables. Then confirmatory factor analysis was used to determine if there is any dormant factor among all the items as it is expected as dominant factor.

The results gained from the estimation of finding item fitness due to CTT assessment showed that only 8 items fitted among twenty (Table 3). The Pearson's coefficient of correlation between each extracted item and the total test score showed that there was significant at $p < .01$, 2-tailed as well as $p < .05$, 2-tailed (Table 4).

The gained answer of question number one "What are the item parameter measurements of participants' responses to linguistics test items based on CTT model?" the obtained results showed that for the CTT model, the item difficulty parameter estimates (*p*-values) ranged from .6 for item 7 to .625 for item 12.As it is shown only items 6 and 7 with *p*-values .8 and .6 were high in difficulty of item.

The discrimination parameter estimates (*a*-value*s*) ranged from 0.00 for item 10 to 0.45 for item 1.The item discrimination can also be referred to as the point biserial correlation, which normally ranges from 0.00 to 1.00 and the higher the value, the more discriminating the item. A highly discriminating item should indicate that students with high test scores responded correctly whereas students with low test scores responded incorrectly. Due to the a-value of CTT 6, 7, 10, 12, and 18 had low discrimination values, then they could be classified as poor items which should be eliminated or completely revised (Table 7).

The answer of question number two "What are the item parameter measurements of participants' responses to linguistics test items based on IRT (2PL) Rasch model?" (Table 8)In case of IRT Rasch model the item difficulty parameter estimates (*p*-values) ranged from -.05 for item 20 to -6.77 for item 8. Test items with high b-values are normally hard items under IRT model; these are the items that low ability examinees are unlikely to answer correctly.

But items with low *b*-values are classified as easy items; these are items that most examinees including the low ability will have at least a moderate chance of answering correctly. all items with a *b*-value greater than 1.0 were classified as difficult items, Out of the 19 items, seven items consists of items 3 (2.41), 5 (-1.20), 6 (-1.64), 13 (1.84), 15 (1.53), 19 (1.13), 20 (6.11) were difficult, specially item 20 was the most difficult one, then to decrease the difficulty of these items they must be revised (Table 8).

The discrimination parameter estimates *a*-values ranged from -.16 for item 10 to -6.11 for item 20.The discrimination value expresses how well an item can differentiate among examinees with different ability levels. Good items usually have discrimination values ranging between 0.5 to 2.0.

High discrimination indicates that higher scoring examinees tend to answer the item correctly, while lower scoring examinees tend to answer the item incorrectly. From among 19 items, 17 items were able to discriminate among the examinees, since their discrimination values were higher than 0.5 (Table 8).

Due to IRT estimation via Rasch model (2PLM), convergence and numerical stability estimation showed the engine status was normal termination, SEM algorithm status was not fully converged and caution was advised, first-order test was convergence criteria satisfied, condition number of information matrix carried 1.38e+002, and second-order test revealed solution is a possible local maximum.

Processing times also showed that *E*-step computation was equal .21, *M*-step computation was .27, *SE* computation equal to 2.14, Goodness-of-fit statistics equal to .16, and totally was 2.79. The answer of question number three "Is there any particular statistical comparability between the CTT and IRT item parameter measurements of participants' responses to linguistics test items?"In addition to correlating the values of CTTzp and IRTzb; CTTza and IRTza, dependent *t-test* was used to find out if the item difficulty and item discrimination parameter estimates by CTT and IRT were statistically significant.

The analysis showed that there was no statistical significant difference between the item difficulty parameter estimates by CTT and IRT ($t19 : -2.000 = p > .05$), and it was also revealed that there was no statistical significant difference between the item discrimination parameter estimates by CTT and IRT($t19=2.000, p > .05$) according to the results shown in Table 10.According to the obtained results, the *p*-values of CTT and *b*-values of IRT were comparable as well as the *a*-values of them. So, as a conclusion, the item parameter estimates of CTT and IRT were used independently for estimating the test

items parameters and showed the comparability of the parameter estimates of the two measurement frameworks.

Discussion and Conclusion

This study was an attempt to assess the comparability test items parameter estimates between Classical test theory (CTT) and Item response theory (IRT) models. The obtained results of the test items parameter estimates of CTT and IRT revealed that the item difficulty and item discrimination values of the two measurement theories had a positive correlation. It was also evident that between CTT and IRT item parameter estimates there was not any significant difference. Consequently, the *p*-values of CTT and *b*-values of IRT were comparable as well as the *a*-values of CTT were comparable with the *a*-values of IRT. Furthermore, the item parameter estimates of CTT and IRT could be used independently to estimate the test items parameters which indicated the comparability of parameter estimates of them. The results of some previous researches performed by Omobola and Adedoyin (2013), ŠpelaProgar and GregorSočan (2008) by supporting MacDonald and Paunonen (2002), also indicated that CTT and IRT measurement theories often produce quite similar results in computing for test items parameter estimates. According to Omobola and Adedoyin (2013),"the terms of ethnics or cultural differences of the examinees which could also affect the assessment of the comparability between CTT and IRT test item parameters estimates are the validity and reliability. The overall results on assessing the comparability of test items parameter estimates between the two measurement theories CTT and IRT could also be affected due to ethnic and cultural background of the examinees, because ethnic, language and cultural background of the examinees are also important factors of psychological and educational testing." It can be concluded from this study that CTT test items parameter estimates are quite similar and comparable to IRT (2PLM) test items parameter estimates as well as the results of estimating the test items parameter revealed that there is a positive correlation between the item difficulty and item discrimination values of the two types of measurement theories. Furthermore, IRT and CTT estimations supported the comparability of these theories in estimating the test items parameters.

REFERENCES

Hambleton RK and Swaminathan H (1985). *Item Response Theory: Principles and Applications* (Kluwer-Nyjhoff) Boston, MA.

Hambleton Ronald K and Swaminathan H (1991). *Fundamentals of Item Response Theory*. Jane Rogers Newbury Park (Sage Publications) CA.

Hambleton Ronald K and Jones Russell W (2005). An NCME instructional module on: comparison of classical test theory and item response theory and their applications to test development. *Journal: Educational Measurement: Issues and Practice* **12**(3) 38-47.

Hmirt Van der Linden WJ and Hambleton RK (1997). *Handbook of Modern Item Response Theory* (Springer Verlag) New York.

Lord FM (1980). *Applications of Item Response Theory to Practical Testing Problems* (Erlbaum) Hillsdale NJ.

Lord FM and Novick MR (1968). *Statistical Theories of Mental Test Scores* (Addison-Welsley Publishing Company) Reading MA.

Novick MR (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology* **3** 1-18.

Omobola O Adedoyin and Adedoyin JA (2013). Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. *Herald Journal of Education and General Studies* **2**(3) 107-114, Available: http://www.heraldjournals.org/hjegs/archive.htm Copyright (c) 2013, *Herald International Research Journals*.

Progar Špela and Sočan Gregor (2008). An empirical comparison of Item Response Theory and Classical Test Theory. *Psihološkaobzorja Horizons of Psychology* **17** 3, 5 24, ©DruštvopsihologovSlovenije, ISSN 1318-187, Znanstveni empiričnoraziskovalniprispevek.

Reeve BB (2002). An Introduction to Modern Measurement Theory. National Cancer Inst.
Schumacker R and Beyerlein ST (2000). Confirmatory factor analysis with different correlation types and estimation methods. *Structural Equation Modeling* **7**(4) 629-636.