Accuracy of Measurement in the Classical and the Modern Test Theory: An Empirical Study on a Children Intelligence Test

Mohammed Mohammed Ateik AL-khadher[1] & Ismael Salameh Albursan[1]

[1] Psychology Department, King Saud University, Riyadh, Saudi Arabia

Correspondence: Ismael Salameh Albursan, Psychology Department, King Saud University, Riyadh, Saudi Arabia. E-mail: ismael_bursan@yahoo.com

Received: November 28, 2015 Accepted: December 22, 2016 Online Published: January 9, 2017

doi:10.5539/ijps.v9n1p71 URL: http://dx.doi.org/10.5539/ijps.v9n1p71

Abstract

The study aimed to compare the accuracy of assessing participants' ability by the significance of standard error of the Classical Test Theory (CTT) and standard error of estimation of the Modern Test Theory (MTT) represented by the Two-Parameter Logistic Model (2PLM). It also aimed to compare item difficulty and arrangement in the two theories using Attriri's Intelligence Scale for Children and a sample of 2674 students from the Republic of Yemen. Descriptive statistics (means and standard deviations), exploratory factor analysis and one-sample t-test were used for statistical treatment of data. Statistical treatment was performed by the IBM SPSS V. 20 and the Bilog-Mg3 programs. It was found that MTT represented by the 2PLM is more accurate than CTT in assessing participants' abilities by standard error. Furthermore, the calibration of items by difficulty and the arrangement of participants' abilities in the two theories proved to be different. Based on the study results, the researcher recommends (a) basing the development of psychological tests on the psychometric characteristics extracted according to MTT, (b) training professionals in measurement and evaluation in the use of analysis programs to extract item and ability parameters according to the different models of MTT (item response theory), and (c) making available the programs needed for the use of MTT in testing, e.g., Xcalibre and Rumm 2030 & Bilog-Mg3.

Keywords: classical test theory, modern test theory, 2PLM, accuracy of measurement

1. Introduction

Most sciences seek to develop objective measures in order to understand, interpret, predict, adjust and control phenomena. Measurements therefore of great importance in any branch of knowledge, provided its procedures are easy to carry out and its results are clear and generalizable. It is noticeable that fundamental developments in psychological and educational measurement methodology have occurred in the past few decades in regard to design, construction or methods of item analysis (Allam, 2001). Tests and measures are necessary for assessment and evaluation. They are tools used to assess development, progress, achievement, underachievement and learning difficulties. They are also used for diagnostic purposes by identifying learners' strengths and weaknesses. Allam (1987) asserts that a test question in human and educational sciences is not limited to its content. It rather is related to responses to the other test items. The reason for this is the overlap between measured traits or abilities. This is characteristic of humanities in comparison with natural sciences that do not have the same degree of overlap. Measurement theories in general assume that there are particular traits or characteristics shared by all individuals who only differ in the degree of these traits and characteristics. Although these traits are invisible ("latent") and cannot be measured directly, they can be inferred and quantified from the observable behavior as represented in the individuals' responses to the items of the test (Wright & Stone, 1979).

Classical Test Theory (CTT) has been widely used in developing tests and interpreting respondents' scores because its assumptions are easy to test and its results are easy to interpret and generalize (Crocker & Algina, 2006). However, this theory has limitations that negatively affect accuracy and objectivity of measurement. This prompted specialists of measurement to seek more accurate and objective measures that do not have the limitations of the CTT-based measures. They wished to take psychological and educational measurement closer to measurement in natural sciences. More specifically, they wished to develop measurements where (a) results are not affected by the items of the tool as long as this tool is suitable for estimating the phenomenon, and (b) the calibration of the tool is based on equal units of measurement that are not affected by the elements at which the phenomenon is estimated (Crocker & Algina, 2006).

Failing to overcome many contemporary psychometric problems and achieve a sufficient level of measurement objectivity, CTT began to lose popularity. This led to the advent of Item Response Theory (IRT). This new theory converts estimates of item parameters to statistics represented on a single comparable scale, so that the respondents "ability and item difficulty are estimated on the same calibration, and respondents" ability is estimated based on their response pattern. This aims to reach invariable item parameters that are independent of the sample and to compute the respondent's latent ability independent of item parameters (item free). This way, respondents' ability and item parameters can be estimated with minimal error (Ismail, 2007). The main idea of these models is linking item characteristics by one parameter or more. Unlike classical models, these models estimate respondent and item parameters with minimal error. Item parameters and respondents' ability are estimated through an iterative process (Embretson & Reise, 2000), without having to use random samples of test items of the measured area and without having to use a very large sample of items representing this area (Hambleton & Swaminathan, 2010).

Despite the theoretical differences between IRT and CTT, concepts of the two theories remain interrelated such that each can compensate for the other's limitations. However, theoretical and practical knowledge of the differences between the two theories in the extraction of respondent and item parameters are still lacking (Silvestre, 2009; Omobola, 2013). Therefore mentioned reasons led to the use of IRT in order to reduce inaccuracy of measurement and lack of objectivity resulting from measurement errors in the raw scores. However, researchers' contributions to eliminate this confusion in IRT are still insignificant. This is the reason why CTT is still dominant and IRT is viewed as a complement, not a competitor (Osterlind, 2006). Negative consequences of traditional measurement procedures can include misleading scores, invalid predictions, uninformed decisions, and judgments of respondents contrary to justice, the ultimate goal of measurement (Aloulila, 2005).

Based on the previous brief account, the present researcher endeavored to compare the accuracy of measurement according to both theories. More specifically, the study addressed the main question of which theory is more accurate and objective in assessing the respondent's ability: CTT or IRT.

1.1 Objectives of the Study

This study aimed to test the accuracy of measurement of both CTT and modern test theory (MTT, an alternative term for IRT) by using the standard error of measurement of CTT and the standard error of estimation of the MTT in regard to items and individuals parameters.

1.2 Statement of the Problem

Researchers frequently depend on CTT in developing tests and analyzing their results. However, the MTT of measurement, as theoretically asserted and as supported by many measurement and evaluation specialists, is preferable for several reasons, including the objectivity of measurement and the calibration of the parameters of items and individuals on the same scale. Another rationale for basing test development and analysis on MTT relates to the standard error of measurement and standard error of estimation which indicate the accuracy of measurement. Thus, this study aimed to empirically compare CTT and MTT for accuracy of measurement using standard error. More specifically, the study addressed the following questions:

1) What are the item difficulty estimates and rankings on the rating calibration according to CTT and the 2-parameter logistic model (2PLM of MTT)?

2) What are the participants' abilities and standard error of measurement according to CTT and the standard error of estimation according to the 2PLM of MTT?

3) Are there significant differences between the mean of standard errors of respondents' abilities between CTT and the 2PLM of MTT?

1.3 Significance of the Study

The study is expected to add to the research that compares the MTT of measurement (IRT) represented by the 2PLM with CTT in regard to accuracy of measurement, as operationalized by its standard error. One aim is to provide practical information which test developers may find useful to achieve measurement objectivity by choosing the theory that achieves the highest degree of accuracy based on the findings. This would be achieved by exploring variation in accuracy of measurement between the two theories in item parameters and respondents' abilities.

1.4 Definition of Terms

1.4.1 Classical Test Theory (CTT)

This is also referred to as the true score theory because it is based on the mathematical model called the true score model. It focuses on traditional reliability theory to assess the strength of the relationship between respondents' observed scores and true scores through indirect measurement.

1.4.2 Two-Parameter Logistic Model (2PLM)

This is one of the models of MTT (also called the Birnbaum model). In this model a discrimination parameter (a_i) is added to the difficulty parameter (b_i) which is represented in the 1PLM assuming absence of guessing (c_i).

1.5 Accuracy of Measurement

Accuracy of measurement in CTT relates to the standard error of measurement for the test as a whole and to test reliability. It refers to the relationship between the observed score and the true score of the respondent' performance on the test (Alsharefean, 2012). In IRT, it refers to the accuracy of item parameter estimates (difficulty and discrimination) and respondents' abilities. It has the highest likelihood that estimation is close to the true value of the parameter or the respondent's ability by choosing unbiased estimators (Alkhader & Aldrabsh, 2014).

1.6 Limitations of the Study

The study was conducted in the second semester of the academic year 2014/2015 on a sample of elementary school graders (from fifth to eighth grades) in the Republic of Yemen represented by ten governorates. Participants' abilities were compared by the significance of the standard error of measurement for CTT and the standard error of estimation for MTT. Item difficulty and arrangement according to the two theories were also explored.

1.7 The Theoretical Framework

1.7.1 Classical Test Theory (CTT)

CTT is the most prevalent theory of measurement. It is based on traditional reliability theory to assess the strength of the relationship between respondent observed scores and true scores (where observed score = true score + error of measurement) through indirect measurement. As a result, this measurement includes measurement error when estimating the respondent's true score (Farrag & A-Sharif, 2013). Charles Spearman (1907-1913) was the first to clarify the basis of CTT in 1900. Thurston (1931), Guilford (1936), Thorndike (1949), Jaliksan (1950), Majnason (1967), and Lord and Novick (1968) also substantially contributed to the development of CTT (Ronald & Cecil, 2012). Spearman found logical mathematical evidence that test scores are measurements that are subject to consistency errors. These errors occur as a result of several factors that we cannot present in distinct sections; rather we can discuss their relative contributions in the fluctuation of reliability scores of the observed scores. That is, the correlation between respondent scores that are subject to error is less than the correlations would be between the true scores in the absence of error.

1.7.2 Item Response Theory (IRT)

IRT represents the modern trend in measurement. This theory with its various models emerged as an extension of CTT, and then generalizability theory. It aims to convert respondents' ability and item parameter estimates to statistics extracted as estimated values of the measured trait (Al-Hakamani, 2007). These measurements have explanatory characteristics that exceed the limits of the total test score to reach an interpretation of respondents' responses on the item measuring this trait (Al-Waliali & Hijazi, 2012). These measurements are represented on a single scale that is comparable, so the respondent's ability and item difficulty are estimated on the same scale (Farrag & Al-Sharif, 2013). This theory and the mathematical models underlying it were based on strong assumptions with which the desired objectivity of measurement can be achieved. These assumptions are: (1) Unidimensionality which means that test items measure a single trait, so the item characteristic curve is one for all group members at a certain level of ability. (2) Local Independence which means that the respondent's response to an item does not positively or negatively affect his/her response to any other item. Local Independencies implicitly achieved if the unidimensionality assumption is achieved (Hambleton & Swaminathan, 2010). (3) Item Characteristic Curve which means that there is a distinctive curve for each test item. This curve represents the relationship between the ability of a respondent and the probability of the correct response to the item, and it varies depending on the model used in the IRT. (4) Freedom from Speediness which means that a respondent's failure to respond correctly to test items is due to limited ability, and not to the effect of speed on the response or his/her inability to complete items because of inadequate test time (Hambleton & Swaminathan, 2010).

1.7.3 The Two-Parameter Logistic Model-2PLM

This model was proposed by the statistician Birnbaum. It assumes that items differ in difficulty (b) and discrimination (a) parameters, i.e., items vary in the slope of their curves and also in their inflection point on the horizontal ability axis. This way, this model does not have the practical difficulty in the one-parameter model by equating the discrimination of all items. This model assumes absence of guessing in constructing items having the same discrimination. Having two parameters (difficulty and discrimination), this model estimates ability based on the general pattern of respondents' correct and incorrect responses to the item. The 2PLM is represented by the following mathematical formula:

$$pi(\theta) = \frac{e^{Di(\theta - bi)}}{1 + e^{Di(\theta - bi)}}$$ (Crocker & Algina, 2006).

1.8 Item Difficulty

In CTT, item difficulty is the percentage of respondents who answer the item correctly and it is symbolized by P. Difficulty ranges from zero to 1: zero for items that all respondents answer incorrectly and 1 for items that all respondents answer correctly. It is computed by the formula: R/N, where R = the number of respondents who answer the item correctly and N = total number of respondents. In MTT represented by the 2PLM, item difficulty is computed by the maximum likelihood method. Item difficulty is identified by its location on the ability scale regardless of the level of discrimination. According to the 2PLM, *b* represents a point on the ability scale when the probability of answering the item correctly = 0.5 (Baker, 2001).

1.9 Individuals' Ability

In CTT, individuals' ability refers to the true score that is estimated from the observed score of the individual's performance on the test. It falls within a certain range referred to as confidence intervals which are determined by extracting the standard error of measurement using test reliability and standard deviation according to the following mathematical formula:

$\mu = \bar{X} \pm Z\sigma_{\bar{X}}$, where ($\mu$) refers to confidence interval boundaries, $\bar{X}$ to the sample mean, Z to the level of trust, $\sigma_{\bar{X}}$ to standard error, and ± to the upper and lower boundaries of the interval. In MTT an individual's ability is assessed from the pattern of his/her responses to test items by using the maximum likelihood method for the possibility of the correct response to every item. The process is repeated until the adjustment becomes small enough to be negligible in estimating the ability of the respondent (Baker, 2001).

1.10 Accuracy of Measurement

Reliability determines the accuracy with which the instrument assesses the measured trait or the characteristics. It means precision in the resulting relationship. It is also an indication of the individual's real performance, whereas the other part refers to performance that is attributable to situational errors. According to the traditional theory, standard error of measurement is computed via reliability, using the following equation: $SEM = \sigma\sqrt{1-R}$ (Abu Hashim, 2010). In MTT it refers to the precision of estimation of item difficulty and individuals' ability parameters, and is characterized with the highest probability that the estimate is close to the true value of the parameter or the true value of the individual's ability by choosing the unbiased estimator and using the standard error of estimation (Al-Khader & Aldrabsah, 2014).

1.11 Calibration Zero

The calibration zero point for item difficulty and individual's ability on the rating scale refers to mean item difficulty by the logit measurement. It is the natural logarithm that increases the likelihood of the individual's responding to items successfully. It indicates item difficulty when the logarithmic preponderant is a fixed amount, which is the natural basis (e = 2.72). The logit unit is the appropriate mathematical unit, as it can be converted into other units that suit different testing conditions (Alanbki, 2009).

1.12 Review of Literature

Some studies compared CTT with the 2PLM. For instance, Stage (2003), experimented with The Swedish Scholastic Aptitude Test (SweSAT). The study aimed to develop a unified higher education admission test by comparing the difficulty parameter in the two theories. Results revealed difference in the estimation of item difficulty between CTT and IRT. Item analysis based on IRT proved to be better than it is in CTT. Similarly, Al-Hakamani (2007) compared CTT and IRT represented in the 2PLM in regard to the estimation of students' ability levels and the stability of the statistical indices of items represented in item difficulty and discrimination. The sample consisted of 3082 male and female students. Results revealed similarity in students' scores estimated according to the two theories in regard to arrangement. However, ability values that were computed according to CTT were relatively different from those that were computed according to 2PLM. Results also showed that item statistical indices that are estimated by the 2PLM were more stable than those estimated using CTT, and that the difficulty index was more stable than the discrimination index in both theories.

Al-Zahrani (2008) conducted a study to explore the effect of sample size and ability spread on the accuracy of estimating the true score by CTT and IRT represented by the 1PLM and the 2PLM. Their test had 60 two-response items. It was found that the 2PLM was better for the estimation of the true score, but significance was not sufficient. The researcher therefore cautioned against over confidence. In brief, research revealed that MTT is better than CTT for the estimation of item and ability parameters, regardless of the caution raised by Al-Zahrani (2008) against overconfidence in MTT.

2. Method

To achieve the aims of the study, the researcher used the descriptive comparative method, which is the best method to identify and compare characteristics according to CTT and the 2PLM to obtain accurate and objective measurement.

2.1 Participants

Participants were 2,647 male and female elementary school students from the fifth to the eighth grades from the main governorates of the Republic of Yemen: Sana'a, Aden, Taiz, Lahij, Ibb, Addali, Al-Hudaydah, Sana'a City, Dhamar and Al-Bayda. Table 1 below presents the distribution of participants according to governorates and gender.

Table 1. The distribution of participants according to governorates and gender

Governorate	Sana'a	Al-Hudaydah	Sana'a City	Addali	Aden	Taiz	Lahij	Ibb	Al-Bayda	Dhamar	Total
Male	51	74	146	98	179	126	23	94	105	433	1329
Female	55	101	247	69	14	73	89	134	143	420	1345
Total	106	175	393	167	193	199	112	228	248	853	2674

2.2 Instrument

The researcher used the Saudi 53-item intelligence test for children (Al-Teriri, 2004), which was adapted to the Yemeni environment by Al-Khader (2012).

2.3 Procedures

The test was graded manually using the correction key developed by the test developer. Item difficulty and ability parameters and standard error of measurement were then extracted according to CTT using the IBM SPSS Statistics 20 program and according to MTT (the 2PLM) using the BILOG-MG3 program.

3. Results

3.1 Testing the Assumptions

Unidimensionality: The researcher used Principal Components Analysis after making sure that conditions for its use were met as shown in Table 2 that shows values of Bartlett criteria (Kaiser-Meyer-Olkin).

Table 2. Bartlett criteria (Kaiser-Meyer-Olkin) for the appropriateness of factor analysis of data

Criteria	Bartlett		
Kaiser-Meyer-Olkin	Chi-Square	df	Sig.
0.878	17438.205	1378	0.000

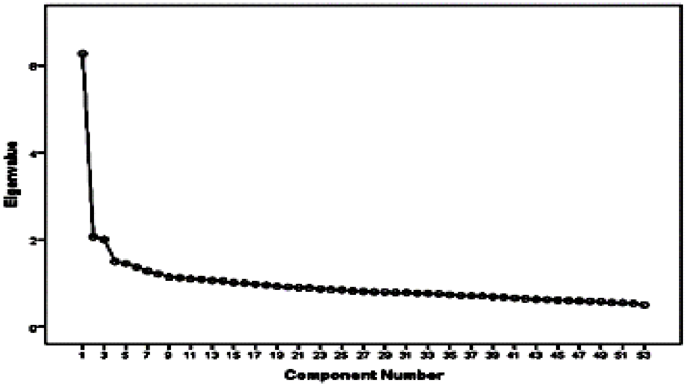


Figure 1. Scree plot of eigenvalues resulting from item analysis

Eigenvalues and percentage of variance for each factor whose Eigenvalue was > 1 were computed. The substantiality of factors was identified at the .30 value as the least value for accepted factor loadings as proposed by Guilford. Eight factors explaining 49% of the variance were extracted. The values of the latent roots of the first and second factors were 6.27 and 2.06 respectively. According to Lord's index (Lord, 1980), the proportion of the first factor to the second factor exceeded 2:1 and the first factor explained 11.84 of the total variance of 49% explained by the 8 factors. Thus, using the rule proposed by Ricki's (listed in Embreston & Riese, 2000), the test proved to be unidimensional as the first factor explained over 20% of the total variance. This was also indicated by the Scree Plot that showed a steep decline between the values of the latent roots of the first and the second factors.

3.1.1 Local Independence

The verification of the unidimensionality assumption is enough to verify the local independence assumption. This was indicated by correlation coefficients among items that did not reach 1, hence proving that no test item was answered based solely on another item (Hambleton & Swaminathan, 2010).

3.1.2 Item Characteristic Curve

This assumption entails the presence of a characteristic curve for each test item. This was verified by the outcome of the statistical analysis done via the BILOG-MG3 program, which included three aspects, one of which is the graphical analysis of items that showed the presence of a characteristic curve for each test item.

3.1.3 Freedom from Speediness

This is another implicit assumption that is verified if the unidimensionality assumption is verified. Factor analysis would reveal the presence of two factors, not one. This indicates freedom from speediness, i.e., speed is not a factor affecting completion of the test. This means that the test is a test of strength, not of speed. In other words, students took sufficient time to answer the test items.

3.2 Appropriateness of Items and Individuals for Analysis according to IRT

The researcher made sure there were no items that all participants answered correctly or that all participants answered incorrectly, which both would be inappropriate to their ability level. Furthermore, participants were excluded who were not appropriate to the calibration process because they failed to answer any item correctly or answered all items correctly (Al-Anbaki, 2009). This way data met the assumptions of IRT. That is, data could be analyzed according to the 2PLM of IRT.

3.3 Study Questions

The identification of the characteristics of test scores gives a brief idea about the test. It is clear from Table 3 that items were appropriate as indicated by the straightness of the correlation coefficients among the variables. This was verified by the mean of items being higher than their standard deviations (Abu-Hashem, 2006).

Table 3. Descriptive statistics of the test

Number	Mean	Standard Deviation	Standard Error
2674	31.63	7.438	3.06

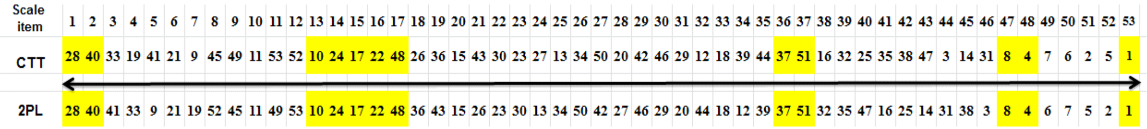
3.3.1 The First Question: What Are Item Difficulty Estimates and Arrangement on the Rating Calibration according to CTT and the 2PLM of MTT?

Item difficulty indices according to CTT ranged from 0.14 to 0.99 with a mean of 0.597. Item 28 was most difficult, while item 1 was the easiest. According to the 2PLM, item difficulty indices ranged from -5.788 to 12.592 logits with a mean of 0.016 logits, indicating that item calibration based on the 2PLM was more accurate, as the difference between the estimates of the difficulty of any two consecutive items is lower than the sum of their standard errors. As the case with analysis based on CTT, item 28 was most difficult, while item 1 was the easiest. According to the two theories, the most difficult items were items 28 and 40 and the least difficult item was item 1. Another nine items (4, 8, 51, 37, 18, 15, 48, 22, 17, 24, 10, 11 and 21) had the same arrangement on the calibration scale of the two theories. The arrangement of these items on the calibration scale was as follows: 48, 47, 37, 36, 33, 20, 17, 16, 15, 14, 13, 10 and 6. The other items were different in the two theories. These findings are consistent with the studies of Stag (2003) and Al-Hakamani (2007). Table 4 and Figure 2 show item difficulty and arrangement according to CTT and MTT.

Table 4. Item difficulty and arrangement on the calibration scale of CTT and MTT

Item Difficulty			**Scale items**		
item	IRT(2PLM)	**CTT**	**item**	IRT(2PLM)	CTT
1	-5.788	0.994	1	28	28
2	-4.36	0.952	2	40	40
3	-2.411	0.801	3	41	33
4	-2.999	0.909	4	33	19
5	-3.182	0.956	5	9	41
6	-3.024	0.95	6	21	21
7	-3.131	0.943	7	19	9
8	-2.457	0.859	8	52	45
9	2.977	0.318	9	45	49
10	0.582	0.441	10	11	11
11	1.487	0.349	11	49	53
12	-0.828	0.641	12	53	52
13	-0.423	0.576	13	10	10
14	-1.708	0.811	14	24	24
15	-0.133	0.519	15	17	17
16	-1.371	0.757	16	22	22
17	0.156	0.478	17	48	48
18	-0.807	0.644	18	36	26
19	2.16	0.298	19	43	36

20	-0.712	0.631	20	15	15
21	2.215	0.311	21	26	43
22	0.121	0.481	22	23	30
23	-0.186	0.549	23	30	23
24	0.28	0.445	24	13	27
25	-1.515	0.766	25	34	13
26	-0.179	0.51	26	50	34
27	-0.621	0.576	27	42	50
28	12.592	0.138	28	27	20
29	-0.707	0.635	29	46	42
30	-0.267	0.535	30	29	46
31	-1.757	0.835	31	20	29
32	-1.225	0.764	32	44	12
33	4.841	0.211	33	18	18
34	-0.471	0.615	34	12	39
35	-1.238	0.776	35	39	44
36	-0.053	0.512	36	37	37
37	-1.051	0.712	37	51	51
38	-1.82	0.787	38	32	16
39	-0.829	0.645	39	35	32
40	12.045	0.157	40	47	25
41	6.684	0.305	41	16	35
42	-0.6	0.632	42	25	38
43	-0.121	0.524	43	14	47
44	-0.748	0.663	44	31	3
45	1.616	0.322	45	38	14
46	-0.696	0.633	46	3	31
47	-1.329	0.799	47	8	8
48	0.01	0.499	48	4	4
49	1	0.328	49	6	7
50	-0.57	0.627	50	7	6
51	-1.183	0.736	51	5	2
52	1.788	0.415	52	2	5
53	0.781	0.356	53	1	1



Scale item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
CTT	28	40	33	19	41	21	9	45	49	11	53	52	10	24	17	22	48	26	36	15	43	30	23	27	13	34	50	20	42	46	29	12	18	39	44	37	51	16	32	25	35	38	47	3	14	31	8	4	7	6	2	5	1
2PL	28	40	41	33	9	21	19	52	45	11	49	53	10	24	17	22	48	36	43	15	26	23	30	13	34	50	42	27	46	29	20	44	18	12	39	37	51	32	35	47	16	25	14	31	38	3	8	4	6	7	5	2	1

Figure 2. The arrangement of items on the calibration scales of CTT and MTT

3.3.2 The Second Question: What Are the Participants' Abilities and Standard Error of Measurement according to CTT and the Standard Error of Estimation according to the 2PLM of MTT?

According to CTT, participants' abilities were estimated based on the total score of participants' performance on the test which ranged from 1 to 50, matching a Z-Score ranging from -4.118 to 2.470 and with a standard error of measurement of 3.06. According to the 2PLM of MTT participants' abilities were estimated in logits by the maximum likelihood method. They ranged from -3.817 to 2.791 logits, where he participant with the number 150 obtained the highest level of ability with a standard error of .619 logits. The participant with the number 1737 achieved the lowest level of ability with a standard error of .361 logits, indicating that ability estimates computed according to the 2PLM and placement on the trait continuum are significantly different from their counterparts computed according to CTT. These results are in line with the studies of Al-Hakamani (2007) and Al-Zahrani (2008). Table 5 below presents these values.

Table 5. Participants' highest and lowest ability estimates and standard errors according to CTT and MTT

Ability	IRT(2PLM)		CTT	
	Ability Logits	standard errors of estimates	Ability Z-Score	standard error of measurement
Highest Ability	2.791	0.619	2.47	3.06
lowest Ability	-3.817	0.361	-4.118	

3.3.3 The Third Question: Are There Significant Differences between the Mean of Standard Errors of Respondents' Abilities between CTT and the 2PLM of MTT?

The One Sample t-test was used to compare the standard error of measurement of participants' abilities according to the CTT (using the IBM SPSS Statistics 20 program) and the 2PLM (using the BILOG-Mg3 program). Results revealed a difference of 2.69 between the means of the standard errors of estimates resulting from the statistical treatment performed according to CTT (= 3.06) and MTT (= 0.372) in favor of MTT. As known, the more accurate estimate is the one that has the least standard error of measurement of ability. This shows that there were significant differences (2-tailed t = -1673.388, df = 2673, p = .000). These results are consistent with the studies of Stag (2003), Al-Zahrani (2008) and AlHakamani (2007). However, they are inconsistent with the study of Osterlind (2006) who concluded that IRT is still complementary to CTT, not a competitor to it.

4. Recommendations

- Basing the development of psychological tests and scales on the psychometric characteristics extracted according to the 2PLM of MTT.
- Training specialists in measurement and evaluation and personnel in the university research centers on the use of analysis software of MTT to extract item and ability parameters according to models of this theory.
- Making available the necessary software required to use MTT in test analysis, e.g., Rumm2030 and Xcalibre & Bilog-Mg3.

Acknowledgments

The authors extend their appreciation to the center of researches in college of education, King Saud University for funding this work.

References

Abu Hashem, A. M. (2010). *Statistical analysis of data using the SPSS program*. Bin Rushd, Riyadh.

Abu-Hashem, A. M. (2006). A comparative study between the classical theory and Rush model in the selection of the items of scale of study entrances to the university students. *Journal of the Faculty of Education, Zagazig University*, *1*, 52.

Al-Anbaki, H. (2009). *Preference in determining the cut-off scores of a criterion-referenced test* (Unpublished Ph. D Dissertation). College of Education, University of Baghdad, EbinRushd.

Al-Hakamani, R. S. (2007). *A comparison between the CTT and the 2PLM in assessing individuals' abilities and stability of test item indices* (Unpublished M. A. Thesis). The College of Education, Sultan Qaboos University.

Al-Khader, M. (2012). *Psychometric characteristics of the intelligent scale for children in the Republic of Yemen* (In Press).

Allam, S. (1987). A comparative critical study of the latent trait models and traditional models in psychological and educational measurement. *The Arabic Journal for the Humanities*, *27*, 16-27.

Allam, S. (2001). *Criterion-referenced diagnostic tests in educational, psychological and training fields*. Dar El-Fekr Al-Arabi, Cairo.

Al-Raheel, R., & Aldarabsah, R. (2014). The effect of the two methods of dealing with missing values and the estimation of ability on the accuracy of estimating item and individual parameters. *The Special International Educational Journal*, *3*, 23-47.

Al-Shareefein, N. K. (2012). The effect of the method of estimating item and ability parameters on item parameter values and the psychometric characteristics of the test in light changing sample size. *The Educational Journal*, *26*(104), 177-238.

AL-Teriri, A. (2004). The intelligence scale for children in the Saudi environment. *Psychological Studies Series*, *8*, 109-139.

Al-Walili, I. (2005). Equivalence of test scores in the light of the classical and modern measurement theories: A comparative psychometric study. *Journal of the Faculty of Education, Benha University*, *63*(15), 51-149.

Al-Walili, I., & Hijazi, A. (2012). The effectiveness analyzing achievement test results according to three-parameter model to predict pupils with learning difficulties in middle school mathematics. *Journal of the Faculty of Education, Benha University*.

AL-Zahrani, B. (2008). *The effect of sample size and ability width on the accuracy of estimating the true by the CTT and unidimensional models of the MTT* (Unpublished Ph. D Dissertation). Faculty of Education, Umm Al Qura University.

Baker, B. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation.

Cecil, R. R., & Ronald, B. L. (2012). *Mastering Modern Psychological Testing Theory & Methods*. Publisher: Pearson.

Crocker, L., & Algina, J. (2006). *Introduction to Classical and Modern Test Theory*. New York: Harcourt Brace Jovanovich College Publishers.

Embreston, S. E., & Riese, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Faraj, M., & AlSharif, K. (2013). Effectiveness of using the latent trait theory and traditional theory in assessing academic achievement in middle school algebra: A comparative study. *The Journal of Imam Muhammad bin Saud Islamic University*, *28*, 89-147.

Hambleton, R., & Swaminathan, H. (2010). *Item Response Theory. Principles and Application*. Boston: Kluwer-Nigh off Publishing.

Ismail, M. A. (2007). *Psychometric characteristics of the mental ability test using Rush among secondary school students* (Unpublished M. A. Thesis). Faculty of Education, University of Zagazig.

Jimelo, L. S. (2009). Item Response Theory and Classical Test Theory: An Empirical Comparison of Item Person Statistics in a Biological Science Test. *Educational and Psychological Assessment*, *1*(1), 19-31.

Omobola, O., & Adedoyin, J. (2013). Assessing the comparability between Classical Test Theory (CTT) and Item Response Theory (IRT) models in estimating test item parameter. *Herald Journal of Education and General Studies*, *2*(3), 107-114.

Osterlind, S. (2006). *Modern measurement: Theory, Principles, and applications of mental appraisal*. Columbus: Pearson.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rash Measurement Chicago*. MESA Press.

Copyrights