THE COLLOQUIUM -A Multi disciplinary Thematic Policy Journal
www.ccsonlinejournals.com
Volume 9 , Number 1 , 2021 Pages 1 - 8
CC-BY-NC-SA 4.0 International Print ISSN : 2971-6624 eISSN: 2971-6632

A COMPARATIVE ANALYSIS OF ITEM DIFFICULTY AND DISCRIMINATION PARAMETERS ESTIMATION IN CLASSICAL TEST AND ITEM RESPONSE THEORIES

UKOFIA, Id-Basil F. & ESSEN, Cyrinus B. Ph.D
Measurement, Evaluation, Research and Statistics Unit,
Department of Educational Psychology, Guidance and Counselling,
Federal College of Education (Technical),Omoku, Rivers State.
Corresponding Email: ibf.ukofia@fcetomoku.edu.ng

Abstract

The global concerns for attaining the best practices in assessment prompted the comparison of classical test theory (CTT) and item response theory (IRT) analysis in psychometrics. These frameworks provide item information needed for the development and use of test items that are capable of estimating examinee's ability and item characteristics. The study compared psychometric properties of item difficulty and discrimination indices in classical test theory and item response theory by the use of BILOG MG3 software. Ex post facto design was adopted for the study. The population for the study consisted of 11,538 candidates' responses of candidates who took Type L 2020 Unified Tertiary Matriculation Examination (UTME) Mathematics paper in Akwa Ibom State, Nigeria. The sample of 5,192 (45%) responses was randomly selected through stratified random sampling technique. Two research questions were raised to guide the study. BILOG-MG3 computer software was used to calibrate the candidates' responses in a 2-Parameter logistic model. The results revealed that some items showed similarities and differences at the various item difficulty indices at both CTT and IRT classification range of values. Also, in CTT item discrimination estimation, point-biserial values and IRT classification range values, showed several degrees of similarities and differences in item calibrations at various indices. However, it was concluded that CTT and IRT are comparable in their classification range values, but IRT provided more reliable information about item classification range values than CTT in psychometrics.

Key words: *Comparability Analysis, Discrimination Parameters, Item Difficulty, Psychometrics, Item Response Theories*

Introduction

The global concerns for attaining the best practices in assessment prompted the comparison of classical test theory (CTT) and item response theory (IRT) analysis in psychometrics. These frameworks provide item information needed for the development and use of test items that are capable of estimating examinee's ability and item characteristics (Essen et al., 2017; Essen & Akpan, 2018; Ryan & Brockmann, 2011). Particular attention should be given to individual items, item characteristics, the probability of answering items correctly, the overall ability of the test taker, and the extent to which an item conforms with the rest of the items in a test (Krishnan, 2013; Rana, 2014). The dominance of Classical Test Theory (CTT) assessment procedures in the educational system in Nigeria, in the 21st century is an issue to be questioned. Many countries of the world are placing the CTT side by side with another complementing test theory known as Item Response Theory (IRT) in the 21st century for best global practices in assessment in educational Measurement. However, one of the reasons for the necessity to adopt these two Measurement theories frameworks is for comparability purposes (Adegoke, 2013; Bichi, 2015; Joshua, 2005).

According to Odili et al., (2015), test developers have shown that test items based on CTT yield results that are slippery and undependable. Therefore, modern testing should be poised at administering test

Cite this article as

Ukofia, I. F., & Essen, C. B. (2021). A comparative analysis of item difficulty and Discrimination parameters estimation in classical test and item response Theories. *THE COLLOQUIUM*, 9(1), 1 -8

items that are objective, informative and challenging to test takers in the educational system. In a comparability analysis study of English Language Paper of Unified Tertiary Matriculation Examination (UTME), Ojerinde (2014) and Alordiah and Ebisine (2016), opined that IRT method is sample independent and no item was removed on the bases of item difficulty. Since CTT is item dependent on the sample used, many items were removed by the classical approach. Furthermore, Morales in Alordiah and Ebisine (2016) carried out a study using multiple Mathematic achievement test of 80 students compared CTT to IRT. The result showed that some items in CTT were found to be bad and were not fitting in IRT models therefore item parameters were inconclusive.Difficulty level of an item determines individual's candidates' chances of giving correct response to such item. Furr emphasized that mathematics item with high difficulty level will not be answered in the same way low difficulty item is answered. Similarly, item with high difficulty level is less endorsed than items with low difficulty level. Thus, the various p-values show items that are considered as: poor or easy ($p > 0.70$), moderately difficult (p $0.31 \leq 0.70$) and difficult ($p \leq 0 . 30$) (Bichi, 2015).Item discrimination indices in CTT statistical procedures is point biserial estimate which is a widely used technique by many researchers (Thompson, 2009; Yaun, et al., 2012; Bichi, 2015). Thompson (2009) held that item discrimination is typically the correlation between item scored dichotomously (0/1) and total test scores, called the item-total correlation. The discrimination level of an item is an indication of how relevant such item is to the ability that is measured in the examination. The item with negative discrimination level is considered not related to the ability measured. Candidates with high ability levels will answer items with high discrimination levels correctly. Therefore, it is imperative that items should have high discrimination levels in a test. The acceptable item discrimination indices are: Very good ($D \geq 0.40$), good ($0.30 \leq 0.39$), marginal ($0.20 \leq 0.29$) and poor ($0.19 \leq$) (Ebel & Frisbie, in Bichi, 2015). The study adopted the item difficulty and discrimination indices values shown in Table 1 and 2 for the analysis. The classifications by Baker (2001) and Baker and Kim (2004) for item difficulty and item discrimination range of values in item response theory was adopted for the study as shown in Table 3 and 4.

Table 1
Classification and Interpretation of Item Difficulty Index in Classical Test Theory

Difficulty index (p)	Interpretation
$P \leq 0.30$	Difficult/hard
$0.31 \leq 0.70$	Moderately difficult
$P > 0.70$	Easy

Source: Henning in Bichi (2015).

Table 2
Classification and Interpretation of Discrimination Coefficient in Classical Test Theory

Item discrimination	Quality of an item	Remarks
$D \geq 0 .40$	Very Good	Item is functioning satisfactorily
$0.30 \leq 0.39$	Good	Item little or no revision is required
$0.20 \leq 0.29$	Marginal	Item may be reviewed
$0.19 \leq$	Poor	Item, should be eliminated

Source: Henning in Bichi (2015)

Table 3 : Item Difficulty Range of Values in Item Response Theory

Ranges	Remarks
-2…….	Very easy
-0.5 to -2	Easy
-0.5 to 0.5	Medium
0 .5 to 2	Hard
2…….	Very hard

Sources: Baker (2001), Baker & Kim (2004)

Table 4 : Item Discrimination Range of Values in Item Response Theory

Ranges	Remarks
0	None
0.01 - 0.34	Very low
0.35 - 0.64	Low
0.65 – 1.34	Moderate
1.35 – 1.69	High
> 1.70	Very Good

Source: Baker (2001), Baker & Kim (2004)

Statement of the Problem

The dominance of CTT assessment procedures in the educational system in Nigeria, in the 21st century is an issue to be questioned. Many countries of the world are placing the CTT side by side with another complementing test theory known as IRT in the 21st century for best global practices in assessment in educational Measurement. However, one of the reasons for the necessity to adopt these two Measurement theory frameworks is for comparability purposes. The presence of problematic items in a test is seen as a threat to reliability and validity, and mars the inference drawn about examinees' ability and proficiency. Further use of such items in subsequent examination without proper item analysis, flaws effective and reliable assessment and becomes a threat to quality educational enhancement. It may cause examining bodies to lack good item banks only to recycle defective instrument for decision-making devoid of reality. One of the National Examination bodies that pioneered the use of IRT in test development and selection is the Joint Admissions and Matriculation Board (JAMB), the body that conducts Unified Tertiary Matriculation Examination (UMTE) for admissions into Nigerian Universities, Polytechnics and Colleges of Education (Ojerinde, 2014). However, the extent to which the use of various software by examination bodies and other psychometrics information is in doubt. The need to examine the comparability analysis of these items in the use of the two theoretical frameworks in the development of items by examination bodies is imperative.

Various computer software used for item development, selection and data analysis in IRT provide both the classical test estimation as well as item response estimation information. BILOG MG, IRTPRO and others, provide comparability information between CTT and IRT in test development and data analysis in psychometrics for reliable and valid estimate in educational assessment, to ensure that comparative information in item selection and analysis is a necessary condition that should be considered in making psychometric decisions. Though, this comparative information seems to be given no serious consideration in test construction and item selection by most of the examination bodies in Nigeria, the need for this comparable information in evaluation and assessment is timely, as better item quality and ability estimates have become acceptable standards of measurement in education and other disciplines in 21st century (NCM in Baker, 2001).

Purpose of the study

The study was carried out to investigate the extent 2020 UTME Mathematics items display comparability at the item parameter indices estimation of item difficulty and discrimination in Classical Test Theory (CTT) and Item Response Theory (IRT) on Mathematics items, using BILOG MG.V3.0 computer software. Specifically, the study compared the:

1. extent 2020 UTME Mathematics items display comparability at the difficulty parameter between CTT and IRT analysis;
2. extent 2020 UTME Mathematics items display comparability at the discrimination parameter between CTT and IRT analysis.

Research questions

The following research questions guided the study:

1. To what extent do 2020 UTME Mathematics items display comparability at the difficulty parameter between CTT and IRT analysis?
2. To what extent do 2020 UTME Mathematics items display comparability at the discrimination parameter between CTT and IRT analysis?

Method

The study adopted the 50 Mathematics items administered by the Joint Admission and Matriculation Board (JAMB) to candidates in 2020 in Akwa Ibom State. The design of the study was ex post facto as the study utilized the data that were marked by the examination body as the researchers were not interested in manipulating the variables. The population for the study consisted of 11,538 candidates' responses who took Type L 2020 Unified Tertiary Matriculation Examination (UTME) Mathematics paper in Akwa Ibom State, Nigeria. The sample of 5,192(45%) responses was randomly selected through stratified random sampling technique. BILOG-MG. V3.0 computer software was used to calibrate the candidates' responses in a 2-Parameter logistic model (difficulty and discrimination). The output was generated at CTT and IRT analysis, making provisions for comparability.

Results

Research Question 1: To what extent do 2020 UTME Mathematics items display comparability at the difficulty parameter between CTT and IRT analysis?

Table 5 : Comparability Analysis of Item Difficulty Indices in Classical Test Theory (CTT and Item Response Theory (IRT)

Classical Test Theory (CTT) Item Response Theory (IRT)

Values	Remarks	Items	Values	Remarks	Items
P>0.70	Easy	2, 3, 4, 5, 7, 8, 9, 10, 12 13, 14, 15, 16, 18, 19, 20, 21, 23 24, 25, 27, 28, 29, 0, 32, 33, 34 36, 37, 38, 39, 41, 42, 43, 44, 45 46, 48, 50, (39 items).	b. -0.5 to -2)	Easy	6, 26,31,40 (4 items)
0.31≤0.70)	Moderate/ Medium	6, 11, 17, 26, 31, 40	-0.5 to 0.5	Moderate/ Medium	No item
(p≤0. 30)	Difficult/ Hard	No Item	.5 to 2	Difficult/ Hard	11, 17, 35, 47

The result revealed a total of 39 items as easy/poor in CTT frameword and 4 items in IRT frameworks. However, CTT indicates 6 items: 6, 26, 31, 35, 40 and 47 at the moderately difficult category (CTT: p 0.31≤0.70). Result further indicates that 4 of the 6 items within the CTT moderate category: 6, 26, 31and 40 are classified by IRT as easy and poor. No item is found at moderate category in IRT. At the difficult/hard index level, IRT locates 4 items: 11, 17, 35 and 47 while no item shows as difficult/hard within CTT framework. While 2 items: 11and 17 found within moderate difficult in CTT is considered as difficult/hard items in IRT framework. IRT shows items: 1, 22 and 49 eliminated in the process of calibration as bad, that did not measure effectively, mathematics ability. By the result, IRT shows accuracy and reliability in item estimate than CTT at all levels of item difficult analysis.

Research question 2: To what extent do 2020 UTME Mathematics items display comparability at discrimination parameter between CTT and IRT analysis?

Table 6 :Comparability Analysis of Item Discrimination Indices in Classical Test Theory (CTT and Item Response Theory (IRT)

Classical Test Theory (CTT)			Item Response Theory (IRT)		
Values	Remarks	Items	Values	Remarks	Items
(rpbis0.19≤)	Poor/low	11, 17, 35, 47,	(a -0.01 to 0.34)	Poor/low	11, 17, 35, 47,
Rpbis0.20≤ 0.29	Marginal/ Moderate	No item	0.65 to 1.34	Marginal/ Moderate	2, 5, 6, 31
Rpbis0.30≤ 0.39	Good/High	2, 31	1.35 to 1.69	Good/High	10, 26, 32, 33, 36, 38, 40, 41, 43, 44, 46, 48, 50
(rpbis.≥ 0.40);	Very good / high	5,6,7,8,9,10, 12, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 25, 26 27, 28, 29, 30, 32,33,34, 36,37, 38,39,40,41, 42, 43,44 45,46, 48 and 50	1.70-2.0	Very good / high	3,4,7,8,9, 12, 13, 14, 15, 16, 18, 19, 20, 21,23, 24, 25, 27, 28, 29, 30, 34, 37, 39, 42, 45

Items at the discrimination parameter between CTT and IRT indicate 4 items: 11, 17, 35, 47 as showing low discrimination values: CTT: (rpbis 0.19≤) and IRT: (a -0.01 to 0.34) indicates similar. At the marginal/ moderate discrimination index level: **CTT:** (rpbis.20 ≤ .29); IRT: (.65 to 1.34), CTT locates no item, while IRT locates 4 items: 2, 5,6,31. The good/high discrimination index level: CTT: (rpbis.30 ≤ .39): IRT: (1.35 to 1.69), CTT indicates items, 2 and 31, while IRT locates 13 items: 10, 26, 33, 36, 38, 40, 41, 43, 44, 46, 48 and 50. However, at the very good/very high discrimination index level: CTT: (rpbis.≥ 0.40); IRT: (1.70-2.0), CTT indicates 39 items: 5,6,7,8,9,10, 12, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 25, 26 27, 28, 29, 30, 32, 33, 34, 36, 37, 38, 39, 40, 41, 42, 43, 44 45, 46, 48 and 50. IRT indicates 26 items. 3, 4, 7, 8, 9, 12, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 25, 27, 28, 29, 30, 34, 37, 39, 42, 45. However, there are issues of comparability as observed in the result in Table 6. Items 2 and 31 located at good discrimination index by CTT: (rpbis0.30 ≤ 0.39): are categorized as having moderate discrimination in IRT: 0.65 to 1.34. Items 5 and 6 located with moderate discrimination index level in IRT are among the items considered as very/high discrimination level. Furthermore, 13 items: 10, 26, 32, 33, 36, 38, 40, 41, 43, 44, 46, 48, 50 located at the good/high category discrimination index in IRT are amongst the items found in the very good/high in CTT discrimination index level.

Cite this article as

Ukofia, I. F., & Essen, C. B. (2021). A comparative analysis of item difficulty and Discrimination parameters estimation in classical test and item response Theories. *THE COLLOQUIUM*, 9(1), 1 -8

Discussion of Findings

The results revealed that the item difficulty parameter indices at their various levels showed some similarities and differences in item calibration. The two frameworks provided comparable item parameter information at their various levels. Though various studies used different methods and software to compare CTT with IRT at difficulty parameter, the IRT classification range values make this study unique in the classification range values of CTT and IRT used. However, the comparability indicates that both frameworks showed some similarities and differences in indicating items at different levels of item difficulty parameter. This study agrees with the empirical studies of Progar et al., (2008), Adegoke (2013), Ojerinde (2013) and Guler et al., (2014) that CTT and IRT are comparable. The study findings also agree with the work of Morales in Alordiah and Ebisine (2016) who used multiple Mathematic achievement test of 80 students to compare CTT to IRT. The result showed that some items in CTT were found to be bad and were not fitting in IRT models. However, this study revealed that though both frameworks are comparable, there are indications of differences and similarities in item identifications at the various range values and levels.

The findings at the item discrimination parameter indices showed similarities and differences in item calibrations at the various discrimination levels in the two frameworks of CTT and IRT. The results agrees with previous studies (Progar et al., 2008; Adegoke, 2013; Ojerinde, 2013; Guler et al., 2014; Alordiah & Ebisine, 2016). The information provided by these findings showed the need to ensure that in educational measurement and assessment, the use of both frameworks in item calibration is for comprehensive and complementary information in decision making in terms of item development and selection for examination purposes. Although the findings of this study point towards similarity between the measurement theories, the most important difference between CTT and IRT is that in CTT, one uses a common estimate of the measurement precision that is assumed to be equal for all individuals irrespective of their attribute levels. In IRT, however, the measurement precision depends on the latent-attribute value. There are other arguments favouring IRT that are worth mentioning. IRT models, including the popular two-parameter logistic and the graded response models (GRMs), take the pattern of item scores into account when inferring latent-attribute scores. Despite theoretical differences between item response theory (IRT) and classical test theory (CTT), there is a lack of empirical knowledge about how, and to what extent, the IRT- and CTT-based item and person statistics behave differently.

Conclusion

From the findings obtained on the comparability of item parameter indices of, item difficulty and discrimination in CTT and IRT, using BILOG-MG V3.0, the two frameworks are very comparable with an indicated similarities and differences in items identified in the calibration process. However, IRT showed superiority in item locations at all levels of indices than CTT. Test items in both measurement frameworks are truly comparable.Therefore placing the CTT side by side with another complementing test theory known as Item Response Theory (IRT) in the 21st century for best global practices in assessment in educational Measurement.

Recommendations

The following recommendations were made:

1. That more studies be carried out to compare item parameters between CTT and IRT in educational assessment to ensure reliability, validity and usability of the items.
2. That other computer software like MULTILOG, IRTPRO, among others, be used to compare the CTT and IRT using different examinations administered or after administration by examination bodies to develop good item banks.
3. That Jamb and other examination bodies in the country should ensure that item parameter analysis is often carried out using the two frameworks of CTT and IRT for the basis of item validation process in item development and selection in standardized test.

Cite this article as

Ukofia, I. F., & Essen, C. B. (2021). A comparative analysis of item difficulty and
Discrimination parameters estimation in classical test and item response
Theories. *THE COLLOQUIUM*, 9(1), 1 -8

References

Adegoke, B.A. (2013). Comparison of item statistics of Physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice*, *4*(22), 87 -96.

Alordiah, C.O., & Ebisine, S.S. (2016). Emerging trend in measurement theories: Do classical test theory and item response theory agree in test item parameter?. *Nigerian Journal of Educational Research and Evaluation*, *15*(1), 56-67

Baker, F. B. (2001). *The basics of item response theory*, 2nd ed. ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S–H. (2004). *Item response theory: Parameter estimation techniques.* 2nd Ed. Marcel Dekker.

Bichi, A.A. (2015). Item analysis using derived science achievement test data. *International Journal of Science and Research (IJSR)*, 4, 1655-1662.

Essen, C. B., Ukofia, I. F., Bassey, B.A., & Idaka, D.O. (2017). Bridging the gap in the current global initiative in validation process in psychometrics: Nigeria perspective. *International Journal of Scientific Research in Education*, *10*(1), 1-11

Essen, C. B. & Akpan, G.S. (2018). Analysis of difficulty and point-biserial correlation indices of 2014 Akwa Ibom state mock multiple choice mathematics test. *International Journal of Education and Evaluation*, *4*(5),1-11

Furr, R. (2007). Item response theoryand Rasch models.Sage Publication Inc. www.Ulb.tu.darmstadt.de/toes/1990.pdf

Guller, N., Uyanik, G. K., & Teker, G. T.(2014). Comparison of classical test theory and item response theory in terms of item parameters. *International Association of Social Science Research*, *2*(1), 1-6. http/iassr.org/journal

Joshua, M. T. (2005). *Fundamental of test and measurement in education*. University of Calabar Press.

Krishnan, V. (2013). The early child development instrument (EDI): An item analysis using classical test theory (CTT) on Alberta's data. htt://www.cup.ualberta.ca/wp-content/uploads/2013/04/ItemAnalysisCTTUPWebsite.pdf.

Odili, J. N. & Osadebe, P. U. (2015). Assessment of stability of item parameter in a mathematics achievement test under the Rach model. Journal of Educational Research and Development (AJERD), 1(2), 1-9

Ojerinde, D. (2014). Innovation in assessment: Jamb experience. A Key Note Address presented at the 16th Annual Conference of the Association of Educational Researchers and Evaluators of Nigeria, University of Calabar

Progar, S., Socan,G., & Slovenija, M.P. (2008). An empirical comparison of item response theory and classical test theory. *Horizon of Psychology*, *17*(8), 5-24.

Ryan, J., & Brockmann, F. (2011). A practitioner's introduction to equating with primers on classical test *theory and item response theory. Handbook* http/www.CCsso.org/Resources/Publication

Thompson, N.A. (2009). Classical item and test analysis with CITAS. *Assessment system Corporation*, 1– 8.http/assess.com/docs/Thomps

Yaun, W., Deng, C., Zhu, H., & Li, J. (2012). The statistical analysis and evaluation of examination results of materials research methods course. *Creative Education*, *3*, 162 http://www.SciRP.org/journal/ce: DOI:10.4236/ce.2012.37B042

Cite this article as

Ukofia, I. F., & Essen, C. B. (2021). A comparative analysis of item difficulty and Discrimination parameters estimation in classical test and item response Theories. *THE COLLOQUIUM*, 9(1), 1 -8