Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Validation of a developed university placement test using classical test theory and Rasch measurement approach

Ado Abdu Bichi*, Rohaya Talib, Noor Azean Atan, Halijah Ibrahim, Sanitah Mohd Yusof

School of Education, Universiti Teknologi Malaysia 81310, Jahor Bahru, Malaysia

ARTICLE INFO

Article history:
Received 15 December 2018
Received in revised form
29 March 2019
Accepted 1 April 2019

Keywords:
Rash model
Classical test theory
Validity
Placement test
Item response theory

ABSTRACT

University entrances examinations are conducted to ensure qualified applicant are placed into appropriate programs of their choices. The outcomes of the test have an important and significant value in taking appropriate decision on the applicant's eligibility, the validity of that examination is paramount to achieving the set goal. The aim of this study is to provide empirical evidence of the construct validity of the newly developed Economics Test using traditional Classical Test Theory and Rasch Measurement Model. The developed Economics Test consists of 70 items after expert judgment and review was administered to 280 students, age 16-20 randomly selected from two public schools in Kano. The study employed a CTT and Rasch model to analyze the data using ITEMAN 4.3 and WINSTEPS 3.72.3 software. The softwares automatically generate the recommended estimate of the parameters to judge the quality of the test items. The results of CTT identified 17 problematic items using difficulty and discriminating index. The results of Rasch showed person statistics (Separation 2.40>2.00 and reliability 0.85>0.80) and item statistics (separation 3.73>3.0 and reliability 0.93>0.8) an excellent person and item reliability. The test measures unidimensional construct supported by the raw variance of 24.9% explained by measures. Investigation of the item person map revealed that the test covered a wide range of the examinees' ability. Overall, using Rasch 10 misfitting construct irrelevant items were identified for deletion. While CTT provides information that is limited to two parameters, the Rasch results provide very detailed information on the quality of the test items. Thus both models can be integrated to generate enough evidence of validity and reliability items in the development of a standardize test.

1. Introduction

Construction and validation of test particularly academic achievement measure includes complex steps, procedures and interrelationship of various ideas and latent variables. Subsequently certain procedures must be followed to develop a test that is firmly identified with the expected outcomes. Two most important steps in test development as spelt out by Haladyna and Downing (2011) are; (i) first, item development which includes content definition, preparation of test specifications, and preparation of the item pool, content validation/experts judgment, pilot testing of the items, data analysis and revision of test items. (ii) Second is item validation through item analysis. All these mentioned processes are closely related with other. Moreover, these processes are carefully executed to ensure only valid and reliable instrument are developed and used to estimate item and person ability. Validity is the foundation upon which all assessment systems are built, whether the assessment tool (Test) is standardized or locally-designed, the aim is to use an instrument that produces true estimate of the examinee ability that could support valid inferences.

The purpose of assessing students learning includes licensing, certification, diagnosis and placement. The entrance examination conducted in universities serves the later purpose (placement) with a view to place qualified applicants into the university's program of their choice. The feedback of university placement examination must have significant values in taking appropriate decision on students' eligibility.

Joint Admissions and Matriculation Board (JAMB) established in 1978 conduct placement or entrance

* Corresponding Author.
Email Address: adospecial@gmail.com (A. A. Bichi)

https://doi.org/10.21833/ijaas.2019.06.004
Corresponding author's ORCID profile:
https://orcid.org/0000-0003-2897-5136
2313-626X/© 2019 The Authors. Published by IASE.

examination called 'Unified Tertiary Matriculation Examination (UTME)' and regulates the admission in all Nigerian universities. All candidates seeking for admissions into undergraduate programs in Nigeria must sit for the UTME. However, the shortcomings (mainly in terms of test administration, scoring and objective decisions) noticed in the process of admitting candidates through UTME led to several calls by stakeholders for an alternative method of admission.

Due to the obvious shortcomings of UTME, the federal government of Nigeria granted power to universities to conduct screening tests 'Post-Unified Tertiary Matriculation Examination (Post-UTME)' in 2005 (Ebiri, 2006). Under this policy it became mandatory for all universities in the country to organize a screening test for prospective candidates after passing their UTME and before offering them a place into their programs. Post-UTME is believed to ensure quality and that, when the best candidates are admitted, the results will also be enhanced which in the long run will lead to the production of better quality graduates from Nigerian universities.

In large scale assessment of this nature, the question of reliability and validity is of great concern. However, the Post-UTME do not follow any professional criteria; because many universities conducts written screening tests consisting of questions that have no any bearing on the candidate's proposed field of study, using unstandardized items which can be more difficult items (Akanwa and Nkwocha, 2015). Similarly, since its' inception to date, there are no sufficient empirical evidences on development, validity and reliability of the Post-UTME despites its' validity and reliability issues. This led to several questions and concern on the Post-UTME validity as such stakeholders suggest among others that, Post-UTME items should be allowed to pass through the processes of standardization and test development and content experts should be involve in developing and validating the Post-UTME items in order to obtain valid and reliable results which will lead to valid interpretations (Ikoghode, 2015).

The challenges face by Nigerian universities today is the need for a standardized test that should be used for the selection of candidate into the undergraduate programs of the universities, a test that would assess the true ability of students and provide valid interpretations with respect to students' eligibility.

This study is conducted to developed and provide a preliminary content and construct validity as well as reliability evidences Economics test for Nigerian Universities. Economics is selected because according to available statistics 53% of the candidates write Economics as compulsory subject for their chosen program in the university (JAMB, 2016). In development and validation of measurement instrument in education and psychology there are two competing frameworks, namely Classical Test Theory (CTT) and Item Response Theory (IRT). The techniques of the two frameworks are applied in instrument development to improve test analysis and refinement procedures (Bichi et al., 2015).

CTT being a traditional approach still attract measurement community in test development and analysis due to its theoretical and practical simplicity. The continuous application of CTT in item analysis is because of its weak assumptions which can easily be met by test data (Champlain, 2010; Hambleton and Jones, 1993). Despite its continuous utilization researchers has question its validity in the present day measurement community (Zaman et al., 2008).

The purpose of this study was to validate the developed Economics Test items for screening applicants into the undergraduate programs of Nigerian Universities using CTT and Rasch Measurement Models. The two models were utilized in order to obtain valid and reliable test items, relevant to measure the true ability of students from traditional and modern measurement perspectives. The Analysis was conducted to determine the appropriate items that satisfied certain criteria for item quality.

2. Classical test theory

Classical Test Theory (CTT) has been widely used for years in determining test item reliability and other characteristics of measurement instruments. CTT is a measurement model in test scores validation that introduces 3 concepts (i) test score (Observed score), (ii) true score, and (iv) error score or random error of measurement. Model has been formulated within this framework (Hambleton and Jones, 1993). The mathematical model is called "Classical Test Model" denoted in Eq. 1.

$$X = T + E \tag{1}$$

This mathematical model is a very simple linear model that links the observable test score(X) to the sum of two unobservable variables, true score (T) and error score (E). Because the true score is not easily observable, instead, the true score must be estimated from the individual's responses on a set of test items. The ability of the students is determined by the number of correct scores obtained by the examinee (Bichi, 2016). Thus the CTT equation is cannot be solve until some simplifying assumptions are made. The major assumptions in CTT are: true scores and error scores are uncorrelated, the average error score of the examinees is zero, and error scores on the parallel tests are uncorrelated (Hambleton and Jones, 1993).

Classical Test Analysis (CTA) utilizes traditional item and sample dependent statistics. These include item difficulty, item discrimination estimates, distractor analyses and a number of related statistics (Bichi, 2016). The analyses in CTT focused on assessment of examinee at the test score level, rather than on the item score level. These analyses include a measure for the reliability (Test level statistics), Difficulty of the Item (Item level statistics) and Discrimination (Item level statistics). Decision on the quality of item according the obtained statistics are taken according to the (Henning, 1987; Ebel and Frisbie, 1991) guideline stated below;

i. Item Difficulty Index = (<0.30) High difficult, (0.31≤0.70) Moderate, (>0.70) Easy items.
ii. Discrimination Index = ≥ 0.40 Excellent, 0.30 ≤ D ≤ 0.39 Good, 0.20 ≤ D ≤ 0.29 Marginal and ≤ 0.19 Poor.

Although CTT developed rapidly with wide application in the measurement community, it has several drawbacks. The weaknesses of CTT are: (i) the estimate of examinee stability depends on the test characteristics (ii) The estimates of item parameter depends on the examinee ability and (iii) The measurement error are only sought for the group not individual student.

Despite the shortcomings attributed to CTT, it was the dominant measurement model until 1953 when Lord published his Doctoral dissertation on Latent Trait Model (Dai-Trang, 2013). Some of these drawbacks in CTT are addressed by the Item Response Theory (IRT). However, is commonly used in test development process because of its simplicity and its test statistics are easy to apply. Whereas CTT approach test outcomes is based on the linear relationship between observed and true score (X = T+ E), in IRT approach, the probability of a response pattern of a test taker as a function of the test taker's ability and the characteristics of the items in a test.)

3. Rasch measurement model

Rasch measurement model was named after Georg Rasch a Danish statistician and mathematician. The Rasch model has two significant properties of internal scaling and invariance these two properties are obtained when the assumption of unidimensionality is met (i.e., when test data fit the model). The model is referred to as a prescriptive model because it prescribes specific conditions for the data to meet. This means that the whole research process, from the very beginning, must be in line with the model's specifications.

One of the basic assumptions of the Rasch measurement model is the unidimensionality: the test should measure one trait at a time. The assumption although theoretically sound, it is practically impossible to construct test which measure only one trait or to prevent the test from the influence of extraneous factors (Baghaei and Amrahi, 2011).

Item Response Theory (IRT) One parameter logistic model (1PL) is widely used as the Rasch model. Rasch followed this existing 1PL. Application of Rasch is considered simple within the IRT Models of two and three Parameter models (2PL, 3PL), because it uses a constant and single parameter scale (D) of 1. Rasch links the opportunities of correct response to each item (P) as a function of examinee ability (θ) [P (θ)] with a constant level of difficulty (b) denoted in an Eq. 2.

$$P(\theta) = \frac{e(\theta - bi)}{1+e(\theta - bi)} \tag{2}$$

Rasch analysis is principally designed to meet the construct validity as described. Item analysis under Rasch focuses on calibration of examinee ability and item difficulty, estimation of model fit, Assessment of unidimensionality as well as distractor analysis. These are the indicator used in measuring the test item quality and relevance to the trait being measured taking into consideration the person ability (Baghaei, 2008). Since its introduction by Georg Rasch in 1960, the application of Rasch in education has led to improvement in learning outcomes and extended to medicine, public health and other disciplines.

4. Methodology

4.1. Participants

The participants in this study comprise two hundred eighty 280 students randomly selected from some senior secondary schools in Kano, Nigeria. The senior secondary school students include male and female, age 16 to 20 years. The stratum (gender) of the students was recognized in the selection of the participants in order to ensure adequate representation of the target population intended for the developed test.

4.2. Instruments

The Economics Test used in this study is a developed 70-item multiple choice. The Test was constructed using the senior secondary schools curriculum in Nigeria. The content of the curriculum with 25 topics was divided into five sections (A-E) in order to ensure content coverage and enhance content validity. The items were adequately distributed using standard test blueprint developed by the researchers and validated by expert. The distribution of the items reflect; Section A (13 items), B (16 items), C (12 items), D (15items) and E (14 items) spread across five domains of Bloom's Taxonomy of Cognitive Objectives. Moreover, a panel of 6 experts was formed to judge/assessed the initial format of the Test from the perspective of economics knowledge and test development criteria. Recommended modifications were made in the instrument based on the expert review, and the produce the first version of 70 test items.

4.3. Administration

The developed Economics Test was administered to the samples by the researchers with the assistance of Economics teachers in sampled schools. Prior to the administration permission was sought and obtained from the appropriate authorities. The purpose of the test was explained to the students and their consents were obtained.

4.4. Data analysis

The data were analyzed using Classical Item Analysis and Rasch Approach; Iteman 4.3 for Classical Analysis and WINSTEPS 3.72.3 for Rasch Analysis. The parameters used to judge the quality of items in CTT were Item Difficulty, Discrimination and Reliability. In Rasch analysis three different stages of estimation were considered, (i) Calibration of examinees' ability and item difficulties (ii) Estimation of fit (iii) Assessment of unidimensionality using Principal Component Analysis (PCA) of Rasch residuals (Bond and Fox, 2007). The relationship between students' ability and item difficulties were presented using person-item maps. The mean square values (MNSQ) and Z standard values (ZSTD) were examined to check the fit statistics.

5. Results and discussions

5.1. CTT analysis results

The result of item analysis using CTT consider three (3) parameters in judging the quality of items to be used in assessing students ability, these are item difficulty (p), Item discrimination (D) and Reliability (r). The results are presented in Table 1 and Table 2.

Summary statistics presented in Table 1 shows that, for the total number of 70 items with 280 examinees, the mean score was 28.67 (SD = 9.67). The mean item difficulty and discrimination are 0.41 and 0.26 respectively. These statistics revealed that, the test has sufficient reliability index according to CTT because, an index of 0.86 which is higher than, the recommended value of 0.70 (Nunnally, 1978).

Table 1: Summary item statistics

Parameter	Value
Number of Items	70
Number of Examinees	280
Reliability (Alpha)	0.861
Mean Scores	28.67
S.D	9.67
Mean P	0.41
Mean r_{pbi}	0.26

The mean item difficulty of 0.41 is within the required standard for moderately difficult item with discrimination index of 0.26 which is not too bad for the entire test (Henning, 1987; Ebel and Frisbie, 1991). The result presented in the Table 2 above indicated that, CTT Item analysis shows that 53 or 75.7% of the items have satisfactory item statistics ($D > 0.19$).

These items satisfied the minimum requirement for inclusion into the final version of test some with minor revision. However, 17 (24.3%) based on the established criteria are recommended to be eliminated from the test having ($D \leq 0.19$). This means that, these 17 defective items are not appropriate and should not be included in the final draft of the test. The internal consistency reliability of the test items was assessed and found to be acceptable with Cronbach's alpha value of 0.862 (Table 1).

5.2. Rasch measurement results

5.2.1. Unidimensionality

To ensure the test is measuring the intended objective, assessing unidimensionality is crucial. To determine the unidimensionality in this study, the PCA of the Rasch residuals was performed. The raw variance explained by measures is 24.9% closely match the expected variance of 24.7%. The raw variance explained by person is 5.8% and that variance explained by items is 19.89%. The results show that, the variance explained of 24.9% is higher than the minimum unidimensionality requirement of 20%, this means that, the unidimensionality is achieved and the test measure a unidimensional constructs.

5.2.2. Person and item reliability

The person reliability and separation indices obtained from the analysis were for 'PERSON RELIABILITY" index is 0.85, and for PERSON'S SEPARATION' value measured was 2.40. This reliability values are considered good, this implies that, the variability in the students' ability in this study is adequate. It is an indication that, the Economics ability of each student was well tested and there are three different groups of students, the low, medium and high achievers (Salleh et al., 2016).

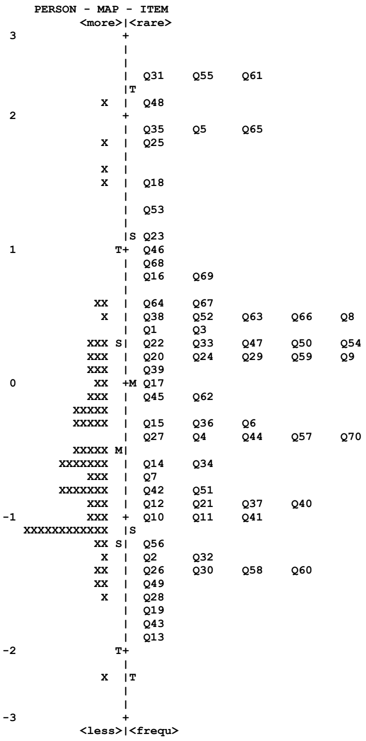
The item reliability and Item separation index were 0.93 and 3.73. These values indicates that, the item reliability in this developed Economics Test is excellent and that, person sample is large enough to confirm the item difficulty hierarchy of the test items.

Table 2: CTT Item analysis chart

Difficulty Index	High Difficult (<0.30)	Moderate (0.31≤0.70)	Easy (>0.70)	Total
Discrimination Index ↓				
Excellent ≥ 0.40	5,8,38,31	3,27,29,45,15		09
Good 0.30 ≤ D ≤ 0.39	18,25,35,66,67,68,63	6,11,21,33,36,41,42,51,54,64,59 62	13,	20
Marginal 0.20 ≤ D ≤ 0.29	16,53	2,4,7,9,10,12,14,17,20,23,24,26,30,34,39,40,44,47,57,22	43,60	24
Poor ≤ 0.19	46,48,52,55,61,65,69	1,32,37,50,56,70	19,28,49,58	17

5.2.3. Respondents-item maps

The relationship between examinees' ability in Economics and the Test items difficulty levels is presented in Person-Item-Map in Fig. 1.



```
    PERSON - MAP - ITEM
           <more>|<rare>
 3               +
                 |
                 |
                 |  Q31     Q55     Q61
                 |T
              X  |  Q48
 2               +
                 |  Q35     Q5      Q65
              X  |  Q25
                 |
              X  |
              X  |  Q18
                 |
                 |  Q53
                 |
                 |S Q23
 1              T+  Q46
                 |  Q68
                 |  Q16     Q69
                 |
             XX  |  Q64     Q67
              X  |  Q38     Q52     Q63     Q66     Q8
                 |  Q1      Q3
            XXX S|  Q22     Q33     Q47     Q50     Q54
            XXX  |  Q20     Q24     Q29     Q59     Q9
            XXX  |  Q39
 0           XX  +M Q17
            XXX  |  Q45     Q62
          XXXXX  |
          XXXXX  |  Q15     Q36     Q6
                 |  Q27     Q4      Q44     Q57     Q70
          XXXXX M|
        XXXXXXX  |  Q14     Q34
            XXX  |  Q7
        XXXXXXX  |  Q42     Q51
            XXX  |  Q12     Q21     Q37     Q40
-1          XXX  +  Q10     Q11     Q41
  XXXXXXXXXXXXX  |S
             XX S|  Q56
              X  |  Q2      Q32
             XX  |  Q26     Q30     Q58     Q60
             XX  |  Q49
              X  |  Q28
                 |  Q19
                 |  Q43
                 |  Q13
-2              T+
                 |
              X  |T
                 |
                 |
-3               +
           <less>|<frequ>
```

Fig. 1: Item-person map

The information from the Map shows the mean value of examinees' ability (M) located on the left side of the map and the mean value of items difficulty (M) placed on the right side of the map. To provide the evidence of representativeness of the test items it can be observe that, the test items are scattered around the mean examinees' ability value. That the item matched with the persons indicating that, the test is targeted for this group of students (Baghaei and Amrahi, 2011). Though the ability of one student was below the difficulty levels of all the items and three (3) items appears to be too difficult for all the test takers. Therefore in order to decide whether to remove or maintain these difficult items and some other that, may display insufficient model fitness there is need to review the model fit of the items to decide whether they indicate a good model fit or not. There is need to investigate the estimation of fit (PTMEA CORR, INFIT MNSQ and OUTFIT MNSQ). Though, there is this little issue, overall the test show acceptable degree of representativeness (Baghaei and Amrahi, 2011).

5.2.4. Model fit statistics

Based on the item map (Fig. 1), there are 3 items which higher than the most able student. The items are Q31, Q55 and Q61. To decide whether omit them from the test or maintain to be use in the next administration, the indicators of fit were investigated i.e., Point Measure Correlation (PTMEA CORR), INFIT Mean Square (INFIT MNSQ) and OUTFIT Mean Square (OUTFIT MNSQ). The investigation was carried out in the entire 70 items to check whether these 3 items and any other violet the standard. Thus, to maintain any item in a test is should satisfy the following conditions as provided by Linacre (2012):

1. PTMEA CORR is positive and not 0 or close to it
2. The INFIT and OUTFIT MNSQ index fall within the acceptable range for Multiple choice Questions, i.e., 0.7≤ MNSQ ≤ 1.3
3. The Z standard (ZSTD) values fall within acceptable range of -2.0≤Z≤2.0

The result shows that, items 31, 55 and 61 Outfit MNSQ are out of the acceptable range and have very low PTMEA CORR close to zero (Linacre, 2012).

Further investigation revealed that, item 48, 49, 50 and 58 also were defective with their outfit MNSQ value exceeding the acceptable range and PTMEA CORR of 0.01, 0.05 and 0.06 close to zero (Table 3). Based on the information all these seven (7) items as indicated should be removed, omitted or revised because of lack of fit to the model.

Table 3: Item statistics (PTMEA CORR, INFIT MNSQ, OUTFIT MNSQ)

Entry	Total	Total	Model		Infit		Outfit		PT-Measure		Exact Match		
Number	Score	Count	Measure	S. E	MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.	Obs.%	Exp.%	Item
31	6	280	2.28	0.44	0.76	-0.61	0.44	-0.4	0.57	0.26	92.5	92.5	Q31
48	7	280	2.1	0.41	1.13	0.5	1.59	1.3	0.06	0.27	91.3	91.2	Q48
49	57	280	-1.51	0.26	1.13	1.2	1.57	2.8	0.02	0.26	71.3	71.7	Q49
50	26	280	0.3	0.25	1.25	2.2	1.31	2.2	0.01	0.33	63.8	71.0	Q50
55	6	280	2.28	0.44	1.13	1.0	1.43	1.0	0.05	0.26	92.5	92.5	Q55
58	56	280	-1.45	0.25	1.12	2.4	1.44	2.4	0.06	0.27	70.0	70.6	Q58
61	6	280	2.28	0.44	0.98	1.8	1.97	1.8	0.17	0.26	92.5	92.5	Q61
MEAN	32.8	280.0	0.0	0.3	1.0	0.0	1.0	0.1			72.8	72.3	
S.D	15.9	0.0	1.1	0.1	0.1	0.7	0.2	0.9			9.0	8.8	

5.2.5. Distractor analysis

In assessing the contribution of distracter to the validating of test items in Rasch, it is expected that, the value for average ability measure should be higher for the correct option and lower for the incorrect options (Linacre, 2012). An asterisk is placed above the average ability measure for correct options that failed to satisfy this condition. This can be observed from items 10, 19, 48, 49, 55, 58 and 70.

Items whose correct options are marked with asterisks should be checked. Conversely, those items manifesting poor model fit and those with higher average ability measures for incorrect options than the correct option should be revised or deleted (Baghaei and Amrahi, 2011). Those which have good fit indices and also the average measures for their wrong options are smaller than the average measure for their correct option are kept. Thus items 10, 19, 48, 49, 55, 58 and 70 having shown the lower means ability measures in the correct options should be deleted, since there is an indication that these distractors do not functions effectively.

6. Misfitting or problematic items by CTT and Rasch

The misfitting items otherwise known as problematic or defective items were identified using the two approaches. The 'problematic' items identified by each framework and the common items identified are presented in Table 4.

Analyses shows that, seventeen (17) items were 'problematic' in that 17 items their discrimination index (DI) is less than 0.19 as stipulated by the highlighted criteria. However, the analyses using Rasch approach ten (10) items were found to have validity issues by misfitting the model thereby classified as 'problematic' as they did not contribute to the validity and reliability of the test. Moreover, the two approaches identified seven (7) common items as 'problematic'.

These results showed that, more items were recommended for deletion by CTT than Rasch this may be connected to the procedures followed by the two frameworks in determining the quality of the test item. While CTT relied on the two parameters of item difficulty and discrimination, Rasch is not limited to item parameters in addition to that person parameter, person reliability, item map, fit statistics and distractors all contribute to the assessment of item misfit, Example item 31 was identified by Rasch and CTT as the difficult item, but CTT classified it as good item because of its discrimination index ignoring its difficulty level.

However, Rasch is able to provide more information based on the ability of the examinees, using Item-person map the item 31 is difficult above the ability of all the examines even the most able student got the item wrong. Similarly, items 10 and 70 were classified as misfitting items by Rasch because the most able students got the items wrong this are some of the additional information given by Rasch that are not feasible with the application of CTT. Looking at the results, some of the items identified as misfit items by CTT are been classified as fit by given more details information based on student's ability. While student ability in CTT is determine based on the raw scores (total) on the exams, the Rasch interpretation of students ability is based on the students responses to difficult and easy items. In CTT students with the same total score will be interpreted as having the same ability. However, in IRT students with same total scores will be interpreted as having different abilities, if one score more on easier item and the other score on difficult items. The student who scores more difficult items will be interpreted as having higher ability. Whereas CTT difficulty values of the item give an indication of how difficult or easy the items are in a test for a group of examinees, Rasch measurement gives a better interpretation of the spread of item difficulty in relation to the examinees' ability levels. Rasch made this feasible through mapping facility (Zubairi and Kassim, 2016).

7. Benchmarking

The major intent of this study was to provide empirical evidence of construct validity as well as reliability of the newly developed Economics Test for Nigerian universities using traditional Classical Test Theory and Rasch Measurement Model (RMM). More importantly was to identify fit/unfit or good or bad items to be maintained or eliminated from the test when the two framework CTT and RMM is used and then to identify the strength and or weakness of each of the two approaches in test development and validation.

The finding of this study shows that, more items were recommended for deletion by CTT than Rasch this may be associated to the techniques followed by the two approaches in determining the feature of the test items. While CTT depend on the two parameters of item difficulty and discrimination, Rasch is not limited to item parameters in addition to that person parameter, person reliability, item map, fit statistics and distractors all contribute to the assessment of item misfit. In the contrary, a study conducted by Abdul-Latif et al. (2016) revealed that, though there was slight difference between item parameter form CTT and RMM, there was no much difference toward the item difficulty provided by CTT and item reliability provided by Rasch Measurement model. The present study is consistent with Petrillo et al. (2015) results were similar when compared between the CTT and RMM, with RMM given more detailed information on how the scale could be improved. The CTT led to the identification of 2 problematic items that threaten the validity and reliability of entire scores of the scale, some sets of item that are redundant and some response that are skewed. Additional RMM identified one item with poor fit and many items that are locally independents. Smiley (2015) the RMM data gives more detailed information that is *sine qua non* for retirement of long term test and development of materials

On the basis of these findings, the selection of a psychometric procedure relies upon numerous elements. Professionals ought to justify their assessment technique and think about the target group. In the event that the test is being constructed for engaging purposes and on a limited spending plan, a superficial examination of the CTT-based psychometric properties might be such is conceivable. In a high-stakes testing like university placement test, however, a careful psychometric evaluation including RMM ought to be considered, with final item level decisions made based on both quantitative and qualitative.

Table 4: Problematic items detected by CTT and Rasch

Model	Number Detected	Items deleted
CTT	17	1,19,28,32,37,46,48,49,50,52,55,56,58,70,61,65,69
Rasch	10	10,19,31,48,49,50,55,58,61,70
Common Items detected	7	19,48,49,50,55,58,61

8. Conclusion

Determining the quality parameters or problematic and good items is an important stage in developing a valid and reliable test items for measuring true ability of students. This study provides item analysis of a developed Economics test using the CTT and Rasch model in order to ascertain its construct validity and reliability evidences. Despite their theoretical as well as methodological differences the two popular frameworks provided a scientific insights on how different test items performed in the developed test by identifying several poor or problematic items using item difficult and discrimination in CTT and Person-Map-Item, Item Fit Statistics (MNSQ, ZSTD and PTMEA CORR) and Item Distracter Analyses in Rasch. The twenty (20) identified 17 CTT and 10 Rasch with 7 common items should be thoroughly investigated with the available information generated from these two frameworks will make the test better. Although, Rasch is theoretically considered to be superior over CTT, several studies found strong relationship between the item parameters obtained using the two approaches.

However, interpretation using Rasch give more detail information on the item structure necessary for valid judgement of student ability and suitability of the items to measure the intended outcome. Similarly, considering the magnitude of the decision to be made from the responses obtained from the administration of higher stake test (such as university placement test), the investigation of test validity should incorporate the two frameworks.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abdul-Latif A, Jalilah I, Amin NFM, and Libunao W (2016). Multiple-choice items analysis using classical test theory and Rasch measurement model. Man in India, 96(1-2): 173-181.

Akanwa UN and Nkwocha PC (2015). Prediction of south eastern Nigerian students' under graduate scores with their UME and Post-UME scores. IOSR Journal of Research and Method in Education, 5(5): 36-39.

Baghaei P (2008). The Rasch model as a construct validation tool. Rasch Measurement Transactions, 22(1): 1145-1146.

Baghaei P and Amrahi N (2011). Validation of a multiple choice English vocabulary test with the Rasch model. Journal of Language Teaching and Research, 2(5): 1052-1060. https://doi.org/10.4304/jltr.2.5.1052-1060

Bichi AA (2016). Classical test theory: Introduction to linear modeling approach to test and item analysis. International Journal for Social Studies, 2(9): 27–33.

Bichi AA, Embong R, and Mamat M (2015). Comparison of classical test theory and item response theory: A review of empirical studies. Australian Journal of Basic and Applied Sciences, 9(7): 549-556.

Bond TG and Fox CM (2007). Applying the Rasch model: Fundamental measurement in the social sciences. Lawrence Erlbaum Associates Publishers, Mahwah, New Jersey, USA.

Champlain DAF (2010). A primer on classical test theory and item response theory for assessments in medical education. Medical Education, 44(1): 109-117. https://doi.org/10.1111/j.1365-2923.2009.03425.x **PMid:20078762**

Dai-Trang L (2013). Applying item response theory modeling in educational research. Ph.D. Dissertation, Iowa State University Ames, Iowa, USA.

Ebel RI and Frisbie DA (1991). Essentials of educational measures. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Ebiri K (2006). Post jamb basis for admission says Obasanjo. The Guardian Newspaper, Kings Place, London, UK.

Haladyna TM and Downing SM (2011). Twelve steps for effective test development. In: Lane S, Raymond MR, and Haladyna TM (Eds.), Handbook of test development: 17-40. Routledge, Abingdon, UK.

Hambleton RK and Jones RW (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12(3): 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

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

Ikoghode A (2015). Post-UTME screening in Nigerian universities: How relevant today?. International Journal of Education and Research, 3(8): 101-116.

JAMB (2016). JAMB statistics: 2016 application and admission. Joint Admissions and Matriculation Board. Nigeria. Available online at: http://www.jambng.com

Linacre JM (2012). A user's guide to Winsteps-Ministep: Rasch-model computer programs. Program manual 3.68.0. Available online at: http://www.winsteps.com/manuals.htm

Petrillo J, Cano SJ, McLeod LD, and Coon CD (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. Value in Health, 18(1): 25-34. https://doi.org/10.1016/j.jval.2014.10.005 **PMid:25595231**

Salleh TS, Bakri N, and Zin ZM (2016). Evaluating a technical university's placement test using the Rasch measurement model. In the AIP Conference Proceedings, AIP Publishing, 1782(1): 1-7.

Smiley J (2015). Classical test theory or Rasch: A personal account from a novice user. SHIKEN, 19(1): 16-29.

Zaman A, Kashmiri AUR, Mubarak M, and Ali A (2008). Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT. In the EDU-COM 2008 International Conference, Edith Cowan University, Perth Western Australia, 1: 281-297.

Zubairi AM and Kassim NLA (2016). Classical and Rasch analyses of dichotomously scored reading comprehension test items. Malaysian Journal of ELT Research, 2(1): 1-20.