Eurasian Journal of Educational Research 84 (2019) 57-70

Eurasian Journal of Educational Research
www.ejer.com.tr

# Rasch-Based Objective Standard Setting for University Placement Test

Ado Abdu BICHI[1], Rohaya TALIB[2], Rahimah EMBONG[3], Hasnah Binti MOHAMED[4], Mohd Sani ISMAIL[5], Abdallah IBRAHIM[6]

ARTICLE INFO

*Article History:*

Received: 1 Jul. 2019

Received in revised form: 19 Sept. 2019

Accepted: 22 Sept. 2019

DOI: 10.14689/ejer.2019.84.3

*Keywords*

*rasch measurement, placement test, cut-scores, objective standard setting*

ABSTRACT

**Purpose:** University placement test is an important admission policy priority in Nigeria, because it serves as a university-based selection criterion for placement of students into undergraduate programs in Nigeria. Although recently attention have been shifted on the call to develop a standard content and standardize the test, yet attention has not been paid on the development of standard setting in which the decisions to select or reject the applicants are made. This study; therefore, investigated the application Rasch-based Objective Standard Setting (OSS) to establish the standards in a university placement test.

**Methods:** To demonstrate the application of OSS, 9 judges were employed for the conduct of standard ratings; the data used for Rasch calibration with WINSTEP were the responses of 600 students on the 60 items validated Economics Placement Test (EPT).

**Findings:** The experts' ratings and Rasch generated logits (item difficulties) were used in quantifying the set of essential items selected. Results of OSS produced the cut-scores of <-0.62 into Basic, -0.62 logits into Proficient and 0.02 logits for Advanced performance levels. The examinees in these categories were, 39% at Basic, 32% at Proficient and 29% at Advanced performance levels.

**Implications for Research and Practice:** The results of this OSS provide performance levels with a clear content related description to informed decision on students' mastery of the content in EPT. It is recommended that, results from the OSS should be compared with other existing IRT-based methods in similar study to ascertain its external validity.



[1] Corresponding Author, School of Education, Universiti Teknologi Malaysia, Johor Bahru, MALAYSIA, e-mail. adospecial@gmail.com, ORCID: https://orcid.org/0000-0003-2897-5136

[2] School of Education, Universiti Teknologi Malaysia, Johor Bahru, e-mail: rohayatalib@utm.my , MALAYSIA, ORCID: https://orcid.org/0000-0002-3937-7193

[3] Faculty of Contemporary Islamic studies, Universiti Sultan Zainal Abidin, MALAYSIA, email: rahimahembong@unisza.edu.my , ORCID: http://orcid.org/0000-0003-0636-8044

[4] School of Education, Universiti Teknologi Malaysia, Johor Bahru, MALAYSIA, ORCID: https://orcid.org/0000-0001-5119-6742

[5] Faculty of Contemporary Islamic studies, Universiti Sultan Zainal Abidin, MALAYSIA, ORCID: https://orcid.org/0000-0003-4608-5639

[6] Centre for Fundamental Studies, Universiti Sultan Zainal Abidin, MALAYSIA, ORCID: https://orcid.org/0000-0003-1698-1192

## Introduction

University placement test serves the purposes of placement of prospective undergraduate into appropriate programmes of universities; therefore, its results or feedback must have important consequences in taking decisions on students' qualification to be placed in any of the university programme (Bichi, Talib, Atan, Ibrahim & Yusof, 2019). The university placement test is conducted to screen students into undergraduate or graduate programmes; therefore, the test has important consequences such as influencing the future achievement of the students. In developing the tests, educational experts should ensure that admission tests are valid and reliable which should be able to predict students' future academic success (Tas & Minaz, 2019; Atkinson, 2001).

All assessment systems are built upon validity, whether the assessment tools (tests) are locally-designed and administered or a standardized test is designed which aims to use a test that produces results to support valid inferences and actions (Atkinson, 2001). Moreover, in educational and psychological testing, the quality of inference generated from the assessment results must be sound and well-structured in principles and empirically verified to withstand systematic criticism (Bichi, et. al, 2019). To produce tests in educational measurement, established criteria and guidelines of valid and reliable test development should be adequately followed. This implies professionalism in both the construction and the use of the test (Sanz & Fernández, 2005).

In Nigeria context, test development remains one of the most tedious aspects in research; there remains less validity and reliability evidence of the university placement test called Post-UTME in Nigeria universities because the Post UTME are developed usually by groups of teachers/lecturers and members of administrative staff of tertiary institutions who incidentally lack the requisites, skills and professional competency to developed and validate placement test of any nature (Akanwa & Nkwocha, 2015; Bichi, 2015; Ikoghode, 2015; Uhunmwuangho & Ogunbadeniyi, 2014) since these tests are constructed for the placement of prospective undergraduate into the Bachelor degree programmes. Similarly, only one study in the literature described the psychometric properties of the Post-UTME Economics i.e Hafiz et al. (2016) and no study described the development and validation of Post-UTME items in Nigeria. Moreover, there is a growing concern over the issue of inappropriate procedure in reporting the students' performances or scores in the test and that is a great mismatch in the appropriateness of the decision taken on students to place them in a particular programme of the university.

In developing standardized tests, especially university placement test, the reliability and validity of the test and reporting of students' performances is essential and inevitable. Thus, in validating test items certain qualities to be considered include item difficulty, item discrimination, quality of the distractors as well as reliability (Barlow, 2014). An evaluation system and assessment programs such as university placement test, reporting student performances is an important concern because a pass-fail decision is taken on students before finally placing prospective

undergraduates in a particular program. Therefore, standards or performance standard is a crucial validity principle and essentially, in high stake assessment methods where the performance standards are used to take critical decision of pass-fail affects especially prospective undergraduate students (Stone et al., 2011).

*Objective of the study*

This study presented the application of Rasch-Based (IRT) Objective Standard Setting Method to establish cut-scores by categorising examinees into Basic, Proficient and Advanced performance levels in a developed university placement test in Nigeria.

*Standard Setting*

Standard setting was used to classify students into different performance standards. Standard setting can be described as a method of generating single cut score, (for example, pass-fail) or multiple cut scores (for example dimension of attainment, excellent, moderate and weak) based on the test requirements or conditions. This cut score work as division of at least two classifications which are necessary for the test (Cizek & Bunch, 2007). Apart from determining the level of students' mastery or achievement, standard setting is a technique applied to obtain cut score which can categorise the examinees into below basic to higher level of performance (Bejar, 2008). As indicated by the report, standard setting is a vital part of test development stages which should include test development professionals, measurement experts and policy makers to ensure that, valid and reliable results are obtained (Bejar, 2008).

There are two classifications of standard setting; norm-referenced (relative) and criterion-referenced (absolute). In norm-referenced classification, standards are determined on the basis of the collective or aggregate performance of the entire group of examinees. Performances are observed between examinee scores as a measure of the whole examinee group and more often used in low-stakes test while criterion referenced standards feature the segment of the test. The standardization depends upon the learning materials and use in high-stakes testing, for example, graduation or final examination.

Basically, there are two popular techniques common among psychometricians and policy makers used in establishing performance standards in a standardized test by classifying the performance levels into (advanced, proficient and basic or below proficient), these are norm-referenced and criterion-referenced standard setting approaches.

*Norm-Referenced Standard Setting (NRS)*

According to Carey and Manwaring (2011) in a norm referenced standard setting procedure, four (4) growth models are applied in determining the examinees' relative performance level, the models include; trajectory, students' growth percentile (SGP), the transition table and projection. Trajectory is the most common among the growth models, because the model requires the identification of the gap between the examinee' current performance level and the already established standard or

proficiency level (Carey & Manwaring, 2011). The second strategy is the student growth model (SGP), this strategy employs norms across particular periods usually years. This model compares the performances of students in similar groups across years to identify whether some level of growth is demonstrated in the current class. The third, which is the transition table, positions or places the examinees into any of the three classifications when their performances are below the proficient level (i.e high minimal, low minimal and weak). Examinees are anticipated to progress to at least to the category higher than their current standing to achieve their annual growth goal (Khatimin, Aziz, Zaharim & Yasin, 2013).

The last and also the most complex model in norm-reference is the projection strategy. This model comprises of two segments. The first is performance standards which is a product of the combination of norm-reference and criterion-reference techniques. It is clear that the standards are certainly not based on content, but alternately is cluttered by referencing to relative or norms. Secondly, it attempts and proposes the prediction of examinees academic progress based on the past record or achievement (Carey & Manwaring, 2011; Silber & Foshay, 2010). The projection model uses advanced level statistics to compare the current examinees with a relatively larger group of similar examinees in the previous years and from different environments to project or predict their success and possibility of attaining the required proficiency. Conclusively, this model considers past performances to predict future success which is believed to be unattainable and unacceptable in the area of high-stake examinations.

*Criterion-Referenced Standard Setting (CRS)*

The Criterion-Referenced Standard Setting can broadly be categorized as traditional and modern. The traditional Criterion-Referenced Standard Setting includes examinees performances and content definition. The examinee performance technique can be seen as an assessment of the content through a generated or quantifiable examinee performance. This method is symbolized in Angoff (1971) which is a popularized model; it can also be seen in other models such as the one offered by Nedelsky (1954), Ebel (1979) and Jaeger (1982). These models require that, an expert or judges in the standard setting exercise evaluate the content given to them in a set of test items and predict the examinee success. Explicitly, the experts or judges are expected to predict the proportion of the examinees who can competently and minimally answer the particular question or item correctly in a given test. At the end of the exercise, a sum or average of their predictions becomes the established performance standard which is commonly come up with after several and consistent iterative sessions where the judges deliberate and reach agreement on the minimally acceptable standard. The Angoff model is considered to be the most popular in this category because it is considered more "sensible, valuable, adequate, and much of the time ideal to different models.

In Modern Criterion-Referenced Standard Setting, Modern Item Response Theory (IRT) techniques such as Rasch approach are popularly been used especially in high-stake testing programmes in the modern assessment setting, this is because the models provide a great avenue and ability in tracking the criterion-referenced performance

growth. When the goal of testing programme is to define the content and apparently a criterion-referenced is content based standard setting method, then defining or describing content through examinees performances using test scores is completely inadequate (Stone et al., 2011; Stone 2009) for the reason that content is only laterally associated to the defined examinee performance and that should not be the basis for its definition. In an effort to solve these critical issues of validity concerns in traditional models, improved modern techniques of establishing standard settings were developed among which are Bookmark Model (Lewis, 1996) and the Objective Standard Setting (OSS) Approach (Stone, 1996). The OSS as modern criterion referenced standard setting model will be the focus in this study.

*Objective Standard Setting Model (OSS)*

Objective standard setting (OSS) defines standards through content directly and not through expectations or likelihood of success (Stone, 2001). A panel of experts or judges are used in the process of setting the standards similar to the traditional approach of standard setting. Conversely, in this technique panels of judges are asked to appraise the contents based on its presentation through the items and determine whether the content is well represented to be considered essential for an examinee to have mastered the content and not the prediction of the likely proportion of successful examinees. Though the entire content is expected to be relatively essential not all content in an assessment or testing program is essential (i.e highly central, core and critical).

Objective standard setting procedure can only be successful with the application of Rasch measurement model. Through Rasch measurement model, raw, deterministic or ordinal test scores are transformed into probabilistic linear-scale identical to ruler with measurement units termed as logits. This transformation holds and improves the full content definition that is necessary in the testing program while creating standard interval level measure. The test scores (crude) failed to address the innate connection among items and its body or with its difficulty level. A raw score of say 75 for instance, may represent 75 correct answers (scores) when applying conventional scoring strategies. Interestingly the Rasch measurement approach tends to address this issue by creating construct ruler that gathers the individual items and individual person parameters along with specific single or more rulers. The test items are arranged in one side and the persons are displayed in another side with easy items and less abled person with less ability at the bottom while most difficult items with high abled persons are placed at the highest point of the ruler (Khatimin, Aziz, Zaharim & Yasin, 2013).

The item measures (logits) represent the construct quantitatively and offer its qualitative expression. Using the logits (item measures) test items are assembled from the less difficult or easiest to the most difficult items on the construct ruler. On the basis of these progresses, this standard setting model has been effectively and constantly utilized in establishing standards setting or performance levels on dichotomously scored tests (MacDougall & Stone, 2015). The ability to apply Rasch measurement models ruler in setting standards lead to the creation of construct within

the OSS, it provides theoretical possibilities to develop multilevel performance standards using the same method. This could make it possible to create two standards or performance levels (advanced and proficient) and apply those levels to a construct ruler.

The OSS was adopted in this study because the need to classify learners into achievement is a frequent and long-standing occurrence internationally (Stone, 1996). The current trends revealed that the development measures are solidly situated in normative gatherings, without reference to explicit substance and eventually utilize normative information which are not criterion referenced. Such self-assertive and norm-referenced standard setting conditions are unsatisfactory and incomprehensible in the realm of high-stakes testing where choices should be legitimately solid and authority of explicit substance is required (Silber & Foshay, 2010). In that capacity, most psychological studies where high-stakes testing is actualized have received modern criterion referencing techniques for standard setting in which OSS approach appears to be more recent, objective and content balanced standard setting technique.

## Method

### *Research Design*

This study intended to establish the standards in a university placement test and to provide evidence of adequacy of the Rasch-based Objective Standard Setting (OSS) in setting standard performance or cut score. Due to the fact that setting cut score in this study was related to the determination and identification of valid and justifiable standard, this was a descriptive survey research (Karasar, 2016).

### *Instrumentation*

The Economics Placement Test (EPT) was a developed 60 items multiple choice test (Bichi, et. al, 2019). The distribution of the 60 items reflects the five (5) Economics sub-dimensions as contained in the National Economics Curriculum of Nigeria (NERDC, 2016) and spread across five domains of Bloom's Taxonomy of Cognitive Objective (revised). The EPT proved to be appropriate to be used with prospective university students as the test possessed adequate content validity (CVR=0.91) with corresponding modified Kappa of 0.864. Equally, the EPT has an excellent person, item as well as adequate internal consistency reliability of 0.87, 0.99 and KR-20 = 0.86 (Linacre, 2019). Similarly, the item measure (infit and outfit) parameters were suitable for all the 60 test items (Bichi, et. al, 2019; Khatimin, et. al., 2013).

### *Sampling and Data Collection Procedure*

To complete the performance standard setting exercise in this study, nine experts were employed to form the panel of judges. All the nine members are experts in Economics Education, assessment and testing because they each taught and participated in the development and administration of the university placement test at universities and colleges in Nigeria. All of the members in the panel hold higher degree (Master and PhD). Nine experts are considered adequate because at least six experts are required to conduct standard setting in OSS (Sondergeld, Stone & Kruse, 2018).

Therefore, on agreement and consent of the experts to participate in this exercise, an OSS item and content description that, contained items and topics, objective for 'item mapping' and performance level descriptions were developed and distributed to each of them. The Judges were asked to review content-balanced item and classify the items as either (i) necessary to demonstrate *minimal proficiency* (ii) necessary to demonstrate *advanced competency* and (iii) not necessary to demonstrate *competency* (Sondergeld, et al., 2018). At the end of the exercise, experts provided their rating on each item individually.

Similarly, to generate item logits for the analysis, the 60 items EPT was administered to 600 prospective university students in Nigeria. The students were through their 3 years senior secondary school and passed through all the Economics contents as spelt in the curriculum, and they were preparing for their final and University entrance examinations that year. Their responses were scored and analysed using Rasch measurement procedures and the item measures; data fit, item and person measures generated were used for selection of essential items.

*Data Analysis*

Experts provided their individual final rating and classification of the items into (i) necessary to demonstrate *minimal proficiency* (ii) necessary to demonstrate *advanced competency* and (iii) not necessary to demonstrate *competency*. Similarly, data from the administration of the EPT was coded and entered in appropriate data sheet for analysis using WINSTEP 3.7.1 (Linacre, 2019) to determine the item measures; data fit, item and person measures as well as item map.

Rasch generated item measures (logits) or item difficulties were used in quantifying the essential items selected. The mean of item measure (logits) linked to each content expert's designated vital or essential test items became the cut score for that expert. The aggregate average score across the experts was used as a cut-score for the test (Sondergeld et. al., 2018). Lastly, considering the fact that all measurement processes were not free from errors, this OSS method took care of it. A sample of content balanced test items below and above criterion within or up to two standard errors of measurements were chosen to produce a standard Rasch construct ruler to determine the ability at *basic*, *proficient* and *advanced* cut scores.

## Results

After the expert judges complete reviews and ratings, the panels of judges classified the items into *"essential"* for content mastery, '*minimal*' proficiency and *not necessary* to demonstrate competency. The practice of OSS began with the selection of essential items identified by the experts within the test.

As required by OSS the previously Rasch calibrated result (measures) was used to quantify the selected items in the two categories. The results in Table 1 showed the mean rating (quantified criterion) associated with each of the panellist's decision on essentials items in logits which became the measure or cut score for that expert judge. An overall average across panel of experts became the final cut score.

The result presented in Table 1 showed that, two standards *advanced* and *proficient* were established by the panelists. The *Advance* standard (0.02 logits) was higher than the *Proficient* standard (-0.62 logits), thus any scores less than -0.62 logits became the *Basic* level or standards. This represented the different achievement scores statistically. Even though some differences existed within the individual expert's ratings between the two levels, the experts set performance standards which represented a strong difference between advanced and proficient standards.

**Table 1**
*Expert Judge Standards (Logits)*

| | Advanced | Proficient |
|---|---|---|
| Expert Judge | (Logits) | (Logits) |
| 1 | 0.20 | -0.11 |
| 2 | -0.12 | -0.36 |
| 3 | 0.16 | -0.96 |
| 4 | 0.09 | -0.56 |
| 5 | -0.33 | -0.88 |
| 6 | 0.05 | -0.80 |
| 7 | -0.21 | -0.76 |
| 8 | 0.16 | -0.70 |
| 9 | 0.16 | -0.43 |
| Mean Logits | 0.02 | -0.62 |

The above information was used with the *Person Statistics* measure order from Rasch analysis to classify the examinees into corresponding performance levels of *Basic* <-0.62 logits, *Proficient* 0.62 logits and *Advanced* 0.02 logits.

Figure 1 depicts the categories in a Rasch construct map, as represented the distribution of the examinees appears to shift upward or higher than the marked proficient level. This is an indication that, the examinees were able to answer moderately difficult and higher difficult items than the less difficult test items.

Based on the logits classifications 39% fell under *Basic* performance level, 32% fell under *Proficient* level and 29% of the examinees fell under *Advanced* performance levels. The final cut scores (Criterion) and the percentage of examinees at each performance levels are presented in Table 2

**Table 2**
*Examinees Category (Cut Scores)*

| **Statistics** | Cut Scores (in Logits) | | |
|---|---|---|---|
| | *Basic* | *Proficient* | *Advanced* |
| Logits | < -0.62 | -0.62 | 0.02 |
| Examinee | 39% (n = 234) | 32% (n = 191) | 29% (n = 175) |

These results means that, 61% of the examinees in this study fell in the category of *Proficient* to *Advanced* level, and that, only 39% were classified as in the *Basic* level, which means that, majority of the examinees were within the required skills (Sondergeld, et al., 2018).

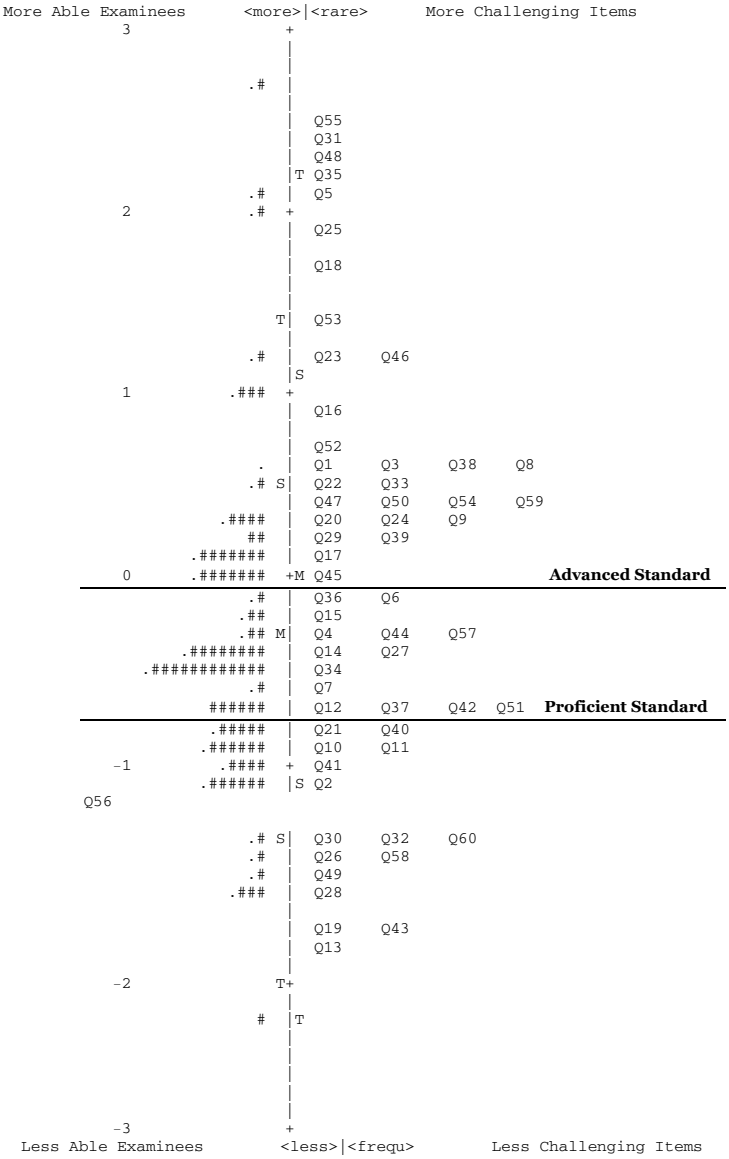


**Figure 1.** *Rasch Person-Map*

## Discussions, Conclusion and Recommendation

This study was conducted to establish cut-scores using IRT-based (*Objective Standard Setting*) method in a university placement test by categorising examinees into Basic, Proficient and Advanced performance levels.

Findings after the rounds of reviews, ratings and discussion regarding the nature of minimal competency revealed that, three standards *Advanced*, *Proficient* and *Basic* were established by the panelists. The Advance standard (0.02 logits) was higher than the Proficient standard (-0.62 logits), thus any scores less than -0.62 logits became the Basic level or standards. The examinees were classified into corresponding performance levels of *Basic* (<-0.62 logits), *Proficient* (0.62 logits) and *Advanced* (0.02 logits). Based on the logits classifications 39% fell under *Basic* performance level, 32% fell under *Proficient* level and 29% of the examinees fell under *Advanced* performance levels. This final cut scores (Criterion) means that, 61% of the examinees in this study fell in the category of *Proficient* to *Advanced* level, and that, only 39% were classified as in the *Basic* level, which means that, majority of the examinees were within the required skills (Sondergeld, et al., 2018).

Based on the finding of this study, Majority of the test takers appeared to be within required levels, with moderately and higher-level ability. The examinee classification showed that, they were able to get the items with moderate and higher difficulty levels right. These findings are consistent with that of (Sondergeld, et al., 2018) while some examines fell within the basic levels, finally, there were test takers in the proficient and advanced levels (n=88.5%, 85%). The largest was the test takers in the advanced classification with n=75 (78.1%). Also, the finding of Khatimin, et. al., (2013) whose findings using OSS revealed that over sixty percent (64%) of the examinees were at the mastery level of the linear algebra and recommended that, academic institutions can decide to adjust the scores to accommodate more students by applying the standard errors (SE).

Similarly, in line with the findings of Stone, Koskey and Sondergeld (2011) whose study included five-year successive investigation by using examination data, examinees were at good performance with a favourable standing, where students who participated in the exercises fell within advanced and proficient standing. In contrary to the finding of this study, Khatimin, Zaharim and Aziz (2014) however, found that, after identifying the mastery levels of -0.08 logits, 74% of the examinees were categorized into performance levels below basic and that, only 26% of the students attained the acceptable mastery levels. This means that, the students in their study did not reach the mastery levels to answer the questions correctly.

Considering the place of standard setting as an important validity principle and more important in high-stakes testing environments, the result of this OSS provided performance level with a clear content related description to informed decisions on students' mastery of the content in a placement test, hence demonstrated effectiveness in designing construct relevant standard and its superiority on establishing standard setting. In a university placement test, reporting students' performance is an important concern because a pass-fail decision is taken on students before finally placed in particular programmes. The cut score proposed in this study can be used in ranking

and selecting the qualified candidates objectively. Therefore, by practitioners utilizing the proposed cut off score in this study, the subjectivity in the selection and placement of prospective undergraduates in Nigerian universities may be reduced or completely eliminated to make the placement of students a transparent process using standardized procedure. It is therefore recommended that to establish more meaningful criterion-referenced standards across the curriculum content being measured, further studies should consider involving or employing more panelists in order to provide more evidence of rating consistency among the panelists. Similarly, result from the Rasch-based Objective Standard Setting (OSS) procedure need to be compared with other existing IRT-based methods in order to ascertain its external validity.

## References

Akanwa, U. N. & Nkwocha, P. C. (2015). Prediction of South Eastern Nigerian students under graduate scores with their UME and Post-UME Scores. *IOSR Journal of Research & Method in Education,* 5(5), PP 36-39

Angoff, W. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp 506-600). Washington, DC: American Council on Education.

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665-676.

Barlow, P. B. (2014). *Development of the Biostatistics and clinical epidemiology skills assessment for medical residents.* Doctoral thesis, University of Tennessee

Bejar, I. I. (2008). Standard setting: What is it? Why is it important. *R& D Connections*, *7*,1-6.

Bichi, A. A. (2015). Analysis of UTME and Post-UTME scores of education students at Northwest University Kano-Nigeria. In *1st International Conference on Education*. Held on 9th-10th April 2015 at Novotel Beijing Xinqiao, Beijing China

Bichi, A. A., Talib, R., Atan, N. A., Ibrahim, H. & Yusof, S. M. (2019). Validation of a developed university placement test using classical test theory and Rasch measurement approach. *International Journal of Advanced and Applied Sciences*, Volume-6, Issue-6, Pp: 22-29

Bichi, A. A., Talib, R., Mohamed, H., Ahamad, J. & Khairuddin, N. (2019). Exploratory sequential design to develop and validate Economics placement test for Nigerian universities. *International Journal of Recent Technology and Engineering (IJRTE)*, Volume-7, Issue-6S5, Pp: 769-772

Carey, K., & Manwaring, R. (2011). Growth Models and Accountability: A recipe for remaking ESEA. Education Sector Reports. *Education Sector*.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications Ltd.

Ebel, R. L. (1979). Essentials of educational measurement, Prentice-Hall. *Englewood Cliffs, NJ*.

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

Karasar, N. (2016). *Scientific research method*. Ankara: Nobel Akademik Yayincilik

Khatimin, N., Aziz, A. A., Zaharim, A., & Yasin, S. H. M. (2013). Development of objective standard setting using Rasch measurement model in Malaysian institution of higher learning. *International Education Studies*, *6*(6), 151-160.

Khatimin, N., Zaharim, A., & Aziz, A. A. (2014, December). Standard setting in students' assessment of higher education institution in Malaysia. In *2014 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)* (pp. 499-504). IEEE.

Lewis, D. M. (1996). Standard setting: A bookmark approach. In *IRT-based standard setting procedures utilizing behavioural anchoring, DR Green (Chair), Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ, 1996*.

Linacre, J. M. (2019). *A User's Guide to WINSTEPS MINISTEP: Rasch-Model Computer Programme*. Online http://www.winsteps.com/winsteps.htm

MacDougall, M., & Stone, G. E. (2015). Fortune-tellers or content specialists: Challenging the standard setting paradigm in medical education programmes. *J Contemp Med Edu*, *3*(3), 135.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*(1), 3-19.

Sanz, I & Fernández, M. (2005). The university entrance exam in Spain: Present situation and possible solutions. *Presented as a poster in the 2005 EALTA Conference*

Silber, K. H., & Foshay, W. R. (Eds.). (2010). *Handbook of improving performance in the workplace* (I). San Francisco, CA: Pfeiffer.

Sondergeld, T. A., Stone, G. E., & Kruse, L. M. (2018). Objective standard setting in educational assessment and decision making. *Educational Policy*, 0895904818802115.

Stone, G. E. (1996). The construction of meaning: Replicating objectively derived criterion-referenced standards. In *annual meeting of the American Educational Research Association. New York, NY*.

Stone, G. E. (2001). Objective standard setting (or truth in advertising). *Journal of applied measurement*, *2*(2), 187-201.

Stone, G. E., Koskey, K. L., & Sondergeld, T. A. (2011). Comparing construct definition in the Angoff and Objective Standard Setting models: Playing in a house of cards without a full deck. *Educational and psychological measurement*, *71*(6), 942-962.

Tas, H., & Minaz, M. B. (2019). An investigation into examination-type preferences of primary school students in relation to various variables. *Eurasian Journal of Educational Research (EJER)*, (81).

Uhunmwuangho, S. O., & Ogunbadeniyi, O. (2014). The university matriculation examination as a predictor of performance in post university matriculation examination: A model for educational development in the 21st century. *African Research Review*, *8*(1), 99-111.