IMPLEMENTATION AND EVALUATION OF STANDARD SETTING: EXPLORING THE UTILITY OF ONE-MEMBER SUB-PANELS USING THE BORDERLINE GROUP METHOD

Inimfon A. Antia & Olusegun G. Ogundare

Abstract

Standard setting has been a concept for consideration in Literature for more than 50 years, and many methods of setting standards have been researched. Senior School Certificate Examinations (SSCE) has also been conducted in Nigeria by both the West African Examinations council (WAEC) and National Examinations Council (NECO) for the certification of candidates for over two decades now. However, formal standard setting procedures seem not to be given much research attention and application by these examining bodies. This study therefore explored the utility of one-member sub-panels for standard setting for a constructed Physics Achievement Test using Borderline Group Method (BGM) in Oyo state, Nigeria. The standard setting study had a sample of 541 Senior Secondary Two (SS2) students, and their Physics teachers purposively sampled from 14 public secondary schools in three local government areas of Oyo state. Four instruments were constructed and used to collect data for the study, including a Physics Achievement Test with reliability (KR_{20} = 0.85). The data obtained were analyzed using descriptive statistics (minimum score, maximum score, mean score, median, standard deviation, Frequency, and Percentages), Item Difficulty, and Independent t-test. Findings of the study revealed that the BGM produced comparable individual cut scores from nine judges' judgement, and a reasonable overall cut score (18) compared with the 50% fixed pass mark of 35. Consequently, the BGM produced a higher Pass/Fail ratio 288/284 (53.23% pass) than the 50% fixed pass mark 26/515 (4.81% pass). The procedures involved in the standard setting exercise, and the training programme were also found to be clear and practicable, by the judges' favourable ratings. Based on the findings, it was recommended that the BGM with one-member sub-panels should be adopted by the examining bodies (WAEC and NECO), and zonal education authorities in setting cut scores for grading their candidates.

Keywords: Standard Setting, Cut Scores, Borderline, One-member Sub-panel, Validity evidence.

Introduction

Assessment is a central process in teaching and learning. The centrality of assessment derives from the need to provide feedback (both to the teacher and to the learners) on behavioural changes or achievement of learners after instruction, or units of instruction in all the three domains of learning (cognitive, affective, and psychomotor) for important decision making. Decisions such as: proceed to next unit/re-teach/repeat unit; pass/fail; promote/repeat; qualified/not qualified; proficient/not proficient admit/reject are some of the decision for which assessment information are collected in schools. Some of the common tools/instruments used to collect assessment data for decision making in the classroom include tests, questionnaires, rating scales, observational schedules among others. Numerical values (results or scores for tests) are usually obtained from these assessment, no matter the instrument used.

Test scores are important pieces of information about test takers (testees). However, test scores in themselves are not meaningful, until they are viewed through the lens of some expectation or standard. To use test scores in decision making, Livingston and Zieky, (1982) recommended two methods; the first involving considering each test takers' test score together with other personal details about the test taker, then applying ones judgment to make a decision. They called this case-by-case method, identifying some advantages and concluding that the major disadvantage is subjectivity. This method offers testees no assurance that they will be treated fairly. The second and better method of decision making involves using a decision rule that will be applied in the same way to all testees. They identified a particularly easy and popularly used decision rule which involves the classifying of testees into what can be called: a higher-scoring group, and a lower-scoring group; giving some examples of test situations where this decision rules is useful to include; promotion, certification, admission/placement. To use any test score with this type of decisions, it is imperative that a test score which will serve to distinguish the two performance groups. This test score is what is known as the “cut score”, which reflect some standard for the particular test and is formally arrived at through some judgmental processes termed standard setting.

Standard setting has been discussed in the literature for more than 50 years and many methods of setting standards have been researched and used (Cizek, 2012; Barman, 2008; Downing, Tekian & Yudkowsky, 2006; Abbott, 2003; Cusimano, 1996; Livingston & Zieky, 1982; Angoff, 1971). Standard setting has been severally defined by different authors, depending on the aspect of the concept considered. From a Practical perspective, standard setting can be defined as the process of establishing cut scores for tests (Cizek & Bunch, 2007). In broad terms, formal standard setting method have been developed to help educators determine which candidates, sitting for a particular test or examination, have performed well enough to pass the assessment and which have not (Schoeman, 2015). Cusimano, (1996) referred to “standard” as a conceptual

boundary (on the true-score scale) between acceptable and non-acceptable performances, while a passing score (cut score) is a particular point (on an observed-score scale) that is used to make decisions about examinees. More explicitly, Reckase's, (2010) definition of standard setting as the label given to the set of activities that are done to identify points on the reporting score scale for a test that represent desired levels of performance was adopted. Further, Hattie and Brown's, (2003) explained that standard setting is a process of obtaining informed judgments from experts who

a. posses adequately knowledge of what the test or assessment for which a cut score to be set requires,
b. have sufficient understanding of the meaning of the varying categories of scores on the scales used to summarize the test takers' performances, and
c. fully comprehend the definitions of achievement that the performance standards that they have been asked to establish carry.

These explanations provide more insights into the practice of standard setting. There are various methods of standard setting, which are generally categorised as examinee-centered, test-centered or a combination of these two approaches (compromise) (Jaeger, 1995). Examinee-centered methods rely upon judgements concerning the test takers. Here, judges segregate the test takers into various performance level (for example non-qualified, qualified and borderline) on the basis of some external benchmark besides the test score (Giraud, Impara and Buckendahl, 2000). And then, the test administrated to the ordered examinees and the cut-score is set dependent on their outcomes on the test (Cizek, 2006). Two popular examinee-centered methods are the borderline-group method and the contrasting group method (Hambleton & Pitoniak, 2006). The examinee centered methods are also known as the norm referenced methods.

Test-centered methods on the other hand are based on judgments about the assessment items. In these methods, the judges take a decisive stand on the level of performance needed to sufficiently satisfy every performance standard (Kane, 1998). This is achieved by deciding on the performance expectation on every test item for hypothetical test takers who are only scarcely satisfying the requirements for a specific performance standard (Hambleton & Pitoniak, 2006). The Angoff (1971) technique and its modified renditions, Ebel, Jaeger's, Nedelsky (1954) methods and the Bookmark method are some of the popular examples of test-centered methods. These methods are also known as criterion referenced.

The methods involving a combination of both test centered and examinee centered methods are often referred to as compromise methods. A compromise method, combining a pre-fixed cut-off score with a relative point of reference, reduces the disadvantages of conventional criterion and norm referenced methods, whilst making optimal use of the advantages (Cohen-Schotanus & Van Der Vleuten, 2010) These methods are thought to produce better cut scores than the norm referenced (examinee

centered) methods and the criterion references (test centered methods). Examples of compromise methods include the Cohen and Hofstee methods

The different methods of standard setting have their peculiarities in various assessment situations, and the choice of a method relies on the merits and demerits of various methods in different contexts (Cizek, 2012). The present study, sought to utilise a formal standard setting approache in the assessment of students' achievement on a researcher made Physics achievement test, with the aim of examining the functionality of a one-member sub-panel design. Considering the practicability of executing a standard setting workshop, under research conditions, the judgement task involved in each of the methods of standard setting were reviewed, and the Borderline Group Method (BGM) was adjudged to be most practicable for the present study in terms of judgements. Also, the conventional 50% fixed pass mark (Adewale & Antia, 2016); will be used to provide additional external evidence for the validation of the process followed in the study, and cut scores set.

The BGM depends on the possibility that the cut score ought to be the score that would be expected from an examinee whose abilities are "on the borderline" — not exactly satisfactory but then not generally lacking (Livingston & Zieky, 1982). In this regard it looks like the methods that are dependent on judgements of test items. In any case, rather than requesting that the judges make surmising about the manner in which a borderline examinee would perform, this method requires the judges to recognize actual examinees as "borderline" in the knowledge and abilities the test measures. The judges not necessarily need to pass judgment on the entirety of the examinees or even a representative sample of them. All they have to do is just recognize the ones who, in their judgment, best fit the description of a borderline test taker. Then the cut score is set at the middle score (the 50th percentile) of this "borderline group".

Notwithstanding the practical simplicity of BGM, the method requires a large sample of examinees performances to give valid and reliable standards and is not recommended for more modest scope testing programs (McKinley, Boulet & Hambleton, 2005). This is because when the number of members in the 'borderline' group is too small, the cut-scores that will be computed may be relatively unstable (Mills, 1995). However, the fundamental bit of leeway of this method remains its simplicity; in terms of usage and training explanations. The major drawback of the BGM is that borderline test takers are oftentimes only but a small portion of the entire examinee group. Consequently, the judges may experience difficulty distinguishing examinees who are truely "borderline".

The BGM can be implemented in the following steps:

1. Selection of judges.
2. Definition of performance levels (that is; adequate, inadequate, and “borderline”)
3. Identifying the “borderline” examinees.

4. Administering the test to get the test scores of the "borderline" examinees.
5. Setting the cut score at the median test score of the borderline group (Livingston & Zieky, 1982).

The cut score so set divides the borderline group into two equal halves, with the use of the median score. The median test score is preferred over the mean score in this method, because the median score is a lot less influenced by outliers (that is few unexpectedly high or low scores). This strengthens the methodology of the BGM, since it can be argued that an examinee who obtains a very high or very low score compared with the scores of other examinees within the borderline group, most likely never actually was supposed to be a member of the borderline group. Therefore, if most of the test scores of the borderline group examinees fall within a narrow range of scores, then the method can be said to be working well. Conversely, if the scores of the borderline group examinees are spread over a considerable range of possible scores, then the method is said not be working well. According to Livingston and Zieky, (1982), some of the factors that can cause the BGM not to work well include:

1. Inclusion of many examinees who are not really borderline in the borderline group, probably because the judges found their skills difficult to judge.
2. Judges making judgments based on something other than what the test measures.
3. Judges showing inconsistencies in their individual standards for judging the test takers.

To avoid these problems in the implementation of the BGM, the proponents of the method provided some suggestion. To avoid the first problem, they suggested that the judges be instructed from the outset, not to include in the borderline group any test takers whose skills they are not familiar with. This is also suggesting that the judges should be selected only from individuals who are familiar with the examinees population, in terms of the knowledge domains to be measured. Also, Livingston & Zieky, (1982) suggested that the second source and third of error can be minimize by giving the judges relevant instructions and getting them to work in agreement with one other, before making their judgments, on a definition of "borderline" knowledge and abilities. This also suggests the standard setting meeting involve lots of reviews and iterative procedures at the stage of understanding the concepts of what borderline entails in the context of the knowledge and abilities to be tested. This would predictably translate to time and cost implications in the process of standard setting, which adds credence to why it is probably jettisoned in the African assessment setting. The present study sought to navigate the spheres of possibilities in the amelioration of such time and cost implications, while upholding the rudiments of the process of standard setting, by exploring the usability of one-member sub-panel with the use of the BGM.

The 50% fixed pass mark is the conventional non-empirical method of setting cut scores in teacher made tests in Nigeria (Adewale & Antia, 2016). In this method, the cut score for a test is preset at 50% of the total obtainable score. This cut score classifies students into two performance levels, which are: credit pass and above, and below credit pass. Here, if the total obtainable score for a given test was 100 marks, a grading system, based on the West African Senior School Certificate Examinations (WASSCE) grading system is employed to categorise students according to their performance. In teacher made tests, raw scores are often distributed along the 9 point grading as: F9 (raw scores 0 – 39); E8 (40 – 45); D7 (45 – 49); C6 (50 – 54); C5 (55 – 59); C4 (60 – 64); B3 (65 – 69); B2 (70 – 74) and A1 (75 – 100). The minimum acceptable performance level in a subject in both WAEC and NECO certificates (which make use of the 9-point grading system) by tertiary institutions in Nigeria is a C6 (credit pass). Hence reducing the 9-point system to a 2-point (pass or fail) decision will imply that passing requires a student to obtain at least a C6, which corresponds to a raw score of 50% in a teacher made testing context. For the present study, the PAT had a total obtainable score of 70 marks hence the 50% fixed pass mark cut score would be a preset score of 35 marks. Therefore, students scoring 35 and above are graded with credit pass and above, while students scoring below 35 are graded with below credit pass.

Several standard setting studies make use of various statistics and indices in three dimensions of validity evidences; procedural evidence, internal evidence and external evidence. For procedural evidence, selection of judges, their profiles/qualifications; training of judges; judges' familiarisation with the process; steps followed through the decision process are often reported as part of the methodology (Elfaki & Salih, 2015; Kollias, 2012; Näsström & Nyström, 2008). For the internal evidence, intrajudge consistency; interjudge consistency; decision consistency and decision accuracy are the prominent indices often presented (Gotzmann, De Champlain, Roy, Brailovsky, Smee, & de Vries, 2014; Kollias, 2012; Berk, 1995). Also for the external evidence, common practice include: comparison of results from two or more standard-setting methods; comparison with grades on other assessments and analysis of consequences of cut scores on the actual scores of examinees (pass/fail rate) (Gotzmann, De Champlain, Roy, Brailovsky, Smee, & de Vries, 2014; Kollias, 2012; Näsström &Nyström, 2008).

To validate the procedures carried out and products obtained, the present study, reported procedural evidence including: selection of judges, judges' profile, training, familiarisation process, decision steps and judges ratings. For the internal evidence, the interjudge consistency index was reported. A common rule-of-thumb applicable to the Angoff method is that low standard deviations between judges indicate high inter-judge consistency and high confidence of the resulting cut-scores (Hambleton & Pitoniak, 2006). Berk, (1995) suggests evaluating intra-panelist (which also was reported) consistency through two sources of evidence, one of which is the consistency of the panelists' ratings to empirical item difficulties and the consistency of individual panelists' ratings across rounds. And to establish external validity, the consequences of

the cut scores set (pass /fail ratios and percentage passes) will be examined alongside standards set by the conventional 50% fixed pass mark. This will serve to reveal the extent to which resulting cut-scores are reasonable which according to Kane, (2001) should be evaluated via impact or pass rates, and alignment with policy considerations.

Considerations for the proposed one-member sub-panel design of the present study using the BGM rely on practicability realities. If a constituted panel of judges are being trained from a central coordinating unit, then in the Nigerian context of education zones; made up of few tens of schools, such coordinating unit can administer the training efficiently working with the panel on individual basis. This will likely increase the possibility of obtaining more valid results. This is because the panellist being subject teachers of sizable number of students (though studies have reported large class sizes) or other personnel working directly with specific students will be more better placed to accurately identify borderline students in the knowledge and abilities of interest than a combined panel of judges working on the overall sample of test takers. Hence, using one-member sub-panels as proposed was hoped to provide a cheaper and more effective variant of the BGM, capable of achieving valid and defensible standard setting for both public examinations such as conducted by WAEC and NECO, and for teacher made examinations that can be adopted by education zones. The study also hope to demonstrate the possibility of researchers' participation in Standard setting, so as to increase focus on, and encourage the utilization of defensible standard setting practice for achievement testing in Africa. The study therefore reported the implementation and evaluation of standard setting for physics achievement test, by exploring the utility of one-member sub-panels made up of secondary school teachers in Oyo state, Nigeria. The following questions guided this study:

Research Questions

1. What are the statistics (minimum scores, maximum scores, mean scores and standard deviation) of students' raw scores for the Physics Achievement Test (PAT)?
2. What are the profiles of the judges involved in the standard setting processes?
3. What are the individual cut scores and overall cut score set for the PAT, using the Borderline Group Method?
4. Is there any significant difference between the individual cut scores set by the selected judges using the BGM and the total item difficulty of the PAT for their borderline students?
5. What are the Pass/Fail Ratios and percentage passes obtained when applying the
   a. Borderline Group Method overall cut scores?
   b. 50% fixed pass mark cut scores?
6. How did the judges' rate the procedures involved in the exercise?

Methods

The present study was a Standard setting study utilising the Borderline Group Method, with one-member sub-panels. The population of the study included all senior secondary school two (SSII) students offering Physics in public secondary schools in Oyo state, Nigeria and their Physics teachers. The study made use of purposive sampling technique for selecting the participating schools, since the study involved a lot of communication between the researcher and the teachers (judges), and there was need for the selected schools to be within a closed circuit. Schools having very small number of science students in SS2 were not selected for the study, and also schools with newly assigned teachers to SS2 Physics were not selected. Three local government areas from Oyo state that are close together (Ibadan north, Ibadan northwest, & Ibadan north east) were purposively selected. Fourteen schools that were as close as possible, after leaving out those that did not meet the criteria for selection were also purposively selected. All the available SS2 students in the selected schools participated in the study. The total sample size was 541 students. The Physics teachers of the selected students in the 14 schools also participated in the study, but only nine of the teachers served as judges at the conclusion of the study.

Four instruments were used in gathering data for this study; Physics Achievement Test (PAT), Training Manual for the Borderline Group Standard Setting Method (TMBGSSM), Borderline Group Sheet (BGS) and Borderline Group Method Evaluation Sheet (BGMES). The PAT was divided into two parts (examinations), similar to the objective and essay parts in the WASSCE Physics. The paper I consisted of I – 50 multiple choice questions, with four options lettered A – D and paper II consisted essay questions having five (5) short answered questions to answer all for 20 marks, making the overall total test maximum obtainable score to be 70 marks. The TMBGSSM consisted of training modules including; introduction to standard setting, information about the students' population, and test, detailed explanation of the Borderline Group Method, and a familiarisation task for participant judges to practice before the actual exercise. The BGS was divided into sections A, and B. Section A contained items requiring information such as local government, type of school, name of school, and judge ID, total number of students taking the test. Section B had a heading; Borderline group students, and contained empty rows for the judge to fill with the test number of borderline students according to their judgment. Finally BGMES was also divided into sections A and B. with section A containing items requiring the following; judge ID, the judges' highest educational qualification, their years of Physics teaching experience, their years of experience as Physics examiners at SSCE level, and their years with present students. And section B containing rating items for the judges to rate the process involved in the method of standard setting they were involved in, and indicate their confidence in the process and in the cut scores set. The reliability of the PAT was found using Kuder-Richardson Formula 20 (KR_{20}) with test data from 75 students. The KR_{20}

reliability coefficient was found to be 0.85 and the face and content validity of the PAT and other instruments were ensured with expert opinion.

The One-Member Sub-Panel Standard Setting Procedure

The procedures followed through in the execution of the one-member sub-panel trial are outlined as follows:

1. **Obtaining permissions for the study:** The researchers after relevant approvals to carry out the study first visited the sampled schools, and teachers, seeking their consent and permission for conducting the study.
2. **Selection and training of judges:** The teachers who consented to take part in the study were trained in the use of the Borderline Group Method through the training manual (individually).
3. **Completion of familiarisation task:** The judges were then allowed to go through and complete the familiarisation task, and their understanding of the process was ensured.
4. **Test administration:** The students were then informed of the test date and asked to prepare. The students were allowed enough time to complete the test, and then the materials were retrieved for marking and further data analyses.
5. **Collection of judgements:** On the test date, the judges arranged the students for the test and assigned them with test numbers. While the tests were on, the judges were given the Borderline Group Sheet (BGS) to record the test numbers of the borderline group students according to their judgments. It was at this stage that five of the judges were dropped, because they provided in their judgements of borderline students, a total number which was less than ten. Not that this was considered incorrect, but they were dropped following the concerns of Mills, (1995) regarding instability of cut scores when there is an extremely small number of students in the borderline group. This made the sample of judges that completed the procedure reduced to nine judges. The overall cut scores set with the nine judges were however used to classify all the students' performance in the study.
6. **Judges evaluation of exercise:** The judges were then required to complete the Borderline Group Method Evaluation Sheet (BGMES), rating their confidence in the process. The judges were later given feedback on their students' achievement on the test. All the 14 judges rated the process because all of them passed through the one-member sub-panel training, and made judgements, even though only nine of the judges' judgements were used for the setting of the cut scores.
7. **Setting cut scores:** From the data generated, individual cut scores (using the BGM), and overall cut scores (using both the BGM and the 50% fixed pass mark) for the PAT were set. For the BGM, the individual cut scores were set at the median raw test score of the borderline group as adjudged by the different judges. For the overall cut

score using the BGM, all the borderline students identified by all the nine judges (a total of 161) were combined as the overall borderline students in the overall study sample. The overall cut score was then set at the median raw test score of these overall borderline students.

Three research assistants were trained and used as invigilators during the test administrations, and data collection which lasted for three weeks. The data collected were analyzed using Descriptive statistics (minimum score, maximum score, mean score, standard deviation) Frequency, and Percentages, Median and item difficulty and paired sample t-test.

Results

Research question 1: What are the statistics (minimum scores, maximum scores, mean scores and standard deviation) and distribution of students' raw scores for the Physics Achievement Test (PAT)?

Table 1: Raw Scores Statistics for the Physics Achievement Test (PAT)

School	N	Min. Score	Max. Score[a]	Mean score	Std Dev.
school 1	75	10	40	22	6.00
school 2	29	14	31	23	4.95
school 3	37	9	25	16	4.24
school 4	65	8	30	18	4.44
school 5	41	5	20	14	3.71
school 6	34	9	29	16	4.21
school 7	40	4	19	13	3.44
school 8	43	10	27	18	4.30
school 9	23	8	25	16	4.31
school 10	28	14	48	33	8.77
school 11	45	7	28	19	4.94
school 12	22	7	19	14	3.56
school 13	19	12	30	21	4.28
school 14	40	17	41	32	5.14
All schools	541	4	48	20	7.44

N = number of students; a = the maximum obtainable score on the test was 70

Result from Table 1 shows the descriptive statistics (minimum scores, maximum scores, mean scores and standard deviation) of the raw scores obtained by students taking the Physics Achievement Test (PAT) in all the sampled schools. The distribution of the results presented shows that the overall minimum score for the test was 4 marks, which was obtained by a student in school 7, while the overall maximum score was 48 marks earned by a student in school 10. With mean scores across the schools ranging from 13 (SD = 3.44) marks (school 7) to 33 (8.77) marks (school 10), which are both below the 50% mark of the test reveals that the PAT was somewhat difficult for the participating student.

Research question 2: What are the profiles of the judges involved in the standard setting exercise?

Table 2: Profiles of Judges Involved in the Standard Setting Exercise

Judge ID	Highest educational qualification	Years of Physics teaching experience	Years of experience as Physics examiner	Years with present students
J1	H.N.D, B. Sc, B, Ed/ equivalent	5 and Above	5 and Above	1 year and above
J2	H.N.D, B. Sc, B, Ed/ equivalent	5 and Above	5 and Above	1 year and above
J4	H.N.D, B. Sc, B, Ed/ equivalent	5 and Above	5 and Above	1 year and above
J5	POSTGRADUATE DEGREE	5 and Above	5 and Above	1 year and above
J6	H.N.D, B. Sc, B, Ed/ equivalent	5 and Above	Below 5	1 year and above
J7	POSTGRADUATE DEGREE	5 and Above	5 and Above	1 year and above
J11	POSTGRADUATE DEGREE	5 and Above	5 and Above	1 year and above
J12	H. ND, B. Sc, B, Ed/ equivalent	Below 5	Below 5	below 1
J13	H. ND, B. Sc, B, Ed/ equivalent	5 and Above	5 and Above	1 year and above

The result from Table 2 shows the profiles (Highest educational qualification, Years of Physics teaching experience, Years of experience as Physics examiner, and Years with present students) of the judges who participated in the standard setting exercise. Six (6) of the judges (J1, J2, J4, J6, J12 & J13) had H. ND, B. Sc, B, Ed and other equivalent degrees as their highest educational qualification. Three (3) judges on the other hand (J5, J7 and J11) postgraduate degrees as their highest educational qualification. Also, eight (8) judges (J1, J2, J4, J5, J6, J7, J11 & J13) had five (5) and above years of Physics teaching experience, whereas only one (1) judge (J12) had below five years of Physics teaching experience. Seven (7) judges (J1, J2, J4, J5, J7, J11, & J13) had five (5) and above years of experience as Physics examiners at O'level, while 2 judges (J6 and J12) had below five years of experience as Physics examiners. Finally, 8 judges (J1, J2, J4, J5, J6, J7, J11 & J13) had been with their present students (the sampled students) for a period of 1 year and above, whereas 1 judges (J12) had been with their present students for a period less than 1 year. These results suggest that the panel of judges selected for the study was sufficiently adequate.

Research Question 3: What are the individual cut scores and overall cut score set for the PAT, using the Borderline Group Method?

Table 3: Individual Cut Scores, Overall Cut Scores, Pass/Fail Ratios for Borderline Group Method and 50% fixed pass mark.

Judge ID	N	N_B	Individual cut score	Pass/Fail Ratio	Percentage pass (%)
J1	75	13	19	51/24	68.00
J2	29	19	23	16/13	55.17
J4	65	30	19	28/37	43.08
J5	41	20	14	24/17	58.54
J6	34	24	16	22/12	64.71
J7	40	16	15	16/24	40.00
J11	45	16	20	19/26	42.22
J12	22	12	14	13/9	59.09
J13	19	11	22	8/11	42.11
Overall	**541**	**161**	**18**	**288/253**	**53.23**

N = Number of students; N_B = Number of identified borderline students

Table 3 presents the borderline students identified by each of the nine judges. As can be seen in the table, judge 4 identified the highest number of borderline student (30) from a sample of 65 students, while judge 13 identified 11 students as borderline line students from a sample of 19 students. The table reveals no fixed proportion of borderline students from the judges working individually. Individual cut scores set by the nine judges are also presented in Table 3. The spread of the individual cut scores spans form 14 marks to 23 marks, which is considerably small. Table 3 also presents the overall cut score set using the BGM (18 marks) from a borderline group of 161 students from the sample of 541 students who took the PAT.

Research Question 4: Is there any significant difference between the individual cut scores set by the selected judges using the BGM and the total item difficulty of the PAT for their borderline students?

Table 4: Paired Sample t-test of difference between Individual Cut Scores and Total Item Difficulty

				Paired Differences					
					95% Confidence Interval of the Difference				
	N	Mean	S.D.	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
BGM_Cut_Score - Total_Item_Diff	9	.00	.50	.17	-.38	.38	.00	8	1.00

Table 4 shows the t-value ($t_{(8)} = 0.00$; $p > 0.05$) of the mean difference between the individual cut scores set for the PAT using the BGM by the nine judges, and their corresponding total item difficulty computed using the borderline group students scores. This implies that there is no difference at all between the means of the two sets of scores.

Research Question 5: What are the Pass/Fail Ratios and percentage passes obtained when applying the

a. Borderline Group Method overall cut scores?

b. 50% fixed pass mark cut scores?

Table 5: Pass/Fail Ratios and Percentage Passes using the BGM and the 50% fixed pass mark

School	N	Borderline Group Method*		50% fixed pass mark	
		Pass/Fail Ratio	Percentage Pass (%)	Pass/Fail Ratio	Percentage Pass (%)
School 1	75	57/18	76.00	3/72	4.00
School 2	29	24/5	82.76	0/29	0.00
School 3	37	11/26	29.73	0/37	0.00
School 4	65	32/33	49.23	0/65	0.00
School 5	41	7/34	17.07	0/41	0.00
School 6	34	11/23	32.35	0/34	0.00
School 7	40	3/37	7.50	0/40	0.00
School 8	43	23/20	53.49	0/43	0.00
School 9	23	7/16	30.43	0/23	0.00
School 10	28	27/1	96.43	14/14	50.00
School 11	45	26/19	57.78	0/45	0.00
School 12	22	4/18	18.18	0/22	0.00
School 13	19	17/2	89.47	0/19	0.00
School 14	40	39/1	97.50	9/31	22.50
Total	**541**	**288/253**	**53.23**	**26/515**	**4.81**

*Overall cut score = 18marks

Results presented in Table 5 shows the consequence data; the Pass/Fail Ratios and percentage passes obtained when applying the BGM overall cut scores, and the 50% fixed pass mark. From the table, a wide gap between the consequences of the BGM cut scores and the 50% fixed pass mark is visible. The BGM overall cut scores is found to consistently cause high percentage passes, while the 50% fixed pass mark was clearly to high for a passing score for the PAT (with only school 10 having a 50% pass and school 14 having 22.50% pass and the rest no passes at all). With the use of the BGM overall cut score (18 marks), the PAT though apparently difficult for the sample of students as seen in Table 1 still resulted in 53.23% pass in the total sample 288 (541) of students. The 50% fixed pass mark on the other hand only allowed for 26 (4.81%) of the students to pass the PAT. The BGM working with one-member sub-panels produced more reasonable cut scores than the conventional 50% fixed pass mark.

Research Question 6: How do the judges' rate the procedures involved in the exercise?

Table 6: Judges Rating of the Procedures Involved in the Standard Setting Exercise

S/N	Items	SD (%)	D (%)	A (%)	SA (%)
1	The training manual provided me with a clear overview of the purpose of the standard setting for the Physics Achievement Test (PAT).	0 (0)	1 (7.1)	7 (50)	6 (42)
2	The training manual answered questions I had about standard setting for the PAT.	0 (0)	1 (7.1)	8 (57.1)	5 (35.7)
3	I have a good understanding of the contents from which items of the PAT are drawn.	0 (0)	1 (7.1)	5 (35.7)	8 (57.1)
4	I have a good understanding of my role in the standard setting activity.	0 (0)	1 (7.1)	7 (50)	6 (42)
5	Reviewing the PAT content helped me to understand the standard setting task.	0 (0)	1 (7.1)	8 (57.1)	5 (35.7)
6	I am well familiar with the students selected for use in this study	0 (0)	0 (0)	6 (42)	8 (57.1)
7	I have a good understanding of the performance level descriptions.	0 (0)	1 (7.1)	7 (50)	6 (42)
8	The training manual is adequately self explanatory	0 (0)	1 (7.1)	8 (57.1)	5 (35.7)
9	I am confident that this training manual will lead to valid cut scores	0 (0)	1 (7.1)	10 (71.4)	3 (21.4)

		No comments	The test is comprehensive	The method is relevant	The exercise requires more time	Experienced teachers should be used for the exercise	The method is useful, only the test should also cover practicals
10	Other comments or suggestions	8 (57.1)	1 (7.1)	2 (14.3)	1 (7.1)	1 (7.1)	1 (7.1)

Total number of judges = 14

Key; SA = **Strongly Agree, A = Agree, D = Disagree, SD = Strongly Disagree**

Results presented in Table 6 show the judges rating of the procedures involved in the standard setting exercise in which they participated. As can be seen from the table, majority of the judges gave favourable responses to all the items, indicating their understanding of the procedures followed, and their confidence in the results the procedures will yield. 8 (57.1) of the judges were completely satisfied with the procedures, and had no further comments or suggestions, while 1 judge stated that "The method is useful, only the test should also cover practicals". Overall, the authors gave favourable ratings of the one-member sub-panel methodology of standard setting.

Discussion of findings

Research Question 1 was aimed at providing statistical information on the overall performance of the students taking the test. From the results, only a few students in the sampled schools were able to obtain such high scores on the test. This suggests that the PAT had high difficulty for the overall sampled students, who were to a large extent homogenous in Physics achievement. This could have been due to the fact that the PAT was structured mostly like the WAEC physics examinations for SS 3 students, whereas the sampled students were only SS2 students. This finding will put in a context the low scores received, and allow for proper interpretations of the students scores using the cut scores set.

Procedural evidence, which is one of the categories of validation evidence given by Kane (2001), includes documentation of the qualification of judges involved in the standard setting exercise. This provides justification for Research Question 2. The findings reveal that the sample of judges used for the study were suitable qualified to participate in setting cut scores Physics Achievement Test. The sample of judges for the study therefore meets Raymond & Reid, (2001) criteria for judge selection, which include "the judges' familiarity with (1) the examinee population; and (2) the intended performance levels to be set." However, little variations are expected from the cut scores as a result of the little variations present in the sample of judges.

Research Question 3 revealed the individual cut scores set for the PAT using the BGM. The essence of the individual cut scores is to measure the consistency of the different judgments provided by the different judges. The overall sampled students are assumed to be similar with only negligible differences, and so the judgments of the judges are expected to be similar if the judges all benefited from the same training. The results presented shows that the spread of the individual cut scores was acceptably low, hence the individual cut scores met Livingstone and Zieky, (1982) condition that, if most of the test scores of the borderline group examinees fall within a narrow range of scores, then the method can be said to be working well. Conversely, if the scores of the borderline group examinees are spread over a considerable range of possible scores, then the method is said not be working well. Also, Research Question 4 provided results that strengthened the internal validity of the individual cut scores. Results show that there was no difference at all between the mean individual cut scores set for the PAT using the BGM by the nine judges, and their corresponding mean total item difficulty computed using the borderline group students' scores. This satisfies Berk, (1995) condition for intra-panelist consistency by ascertaining the consistency of the panelists' ratings to empirical item difficulties

Since the comparability of individual cut scores set by the different judges for the PAT is ascertained, then the overall cut score set using the BGM were therefore considered. Results presented showed that the overall cut score set using the BGM (18 marks) was

reasonable, and so was generalisable over the entire students' population. This is in accordance with Hambleton's, (2001) submission that "if it cannot be demonstrated that similar performance standards would result with a second panel, the generalisability of the performance standards is limited, and the validity of the performance standards is significantly reduced". The individual and overall cut scores set using the BGM were considerably low, relative to the 50% fixed pass mark. These low cut scores can be tied to the judgments made by the judges about the borderline students. This is also an extension of their understanding of the performance level descriptions. These low cut scores are however are in conformity with the study of , Parmar, Shah and Parmar, (2014), who compared the Modified Angoff Method with the Hofstee method in setting standards for certificate and licensure Forensic Medicine examinations at Government Medical College Bhavnagar, India. They found that that Cut-off score of all 5 years by Modified Angoff Method was set below 50%.

To provide external validity for the overall cut score set using the BGM, Research Question 5 explored the comparability of the consequences of the BGM overall cut score (pass/fail ratio and percentage pass) and the 50% fixed pass mark. Findings revealed that the BGM working with one-member sub-panels produced more reasonable cut scores than the conventional 50% fixed pass mark. This finding was in line with Kane, (2001) external validity requirement that external validity evidence should be evaluated via impact or pass rates, and alignment with policy considerations.

Finally, the evaluation of the process followed through in implementing the one-member sub-panel standard setting study was the aim of Research Question 6. Findings show that the judges rated the procedures well in the light of the items found on the Borderline Group Evaluation Sheet. This positive rating of the procedures by the judges and high judges' confidence in the procedures, provide judges' feedback; a procedural validation for the process involved in the standard setting study. It demonstrates uniformity, and clarity in the training process, and practicability of the requirements from the judges. This is in tandem with Kane's, (2001) discussion on the importance of panellist feedback, where he argued that this type of evidence can be collected easily and often throughout the standard setting. For example, surveys can be completed after orientation, training, rounds in the standard setting, and after completing the full standard setting meeting. The useful comments made by some of the judges also declare their understanding and interest in the study. On the other hand those comments suggests acceptance of the methods, and identifies the need for further attention and development of the methods in order that they could be useful in the actual exercise of standard settings for SSCE, and other similar certificate examinations.

Conclusion

Appropriate, empirically backed, and objective standard setting is fundamental to good assessment in education. When measurement results are to be used in assessments involving high stakes decisions such as licensure or certification, defensibility of standards used in such decisions should be paramount. The one-member sub-panel design using the BGM explored in this study yielded comparable and reasonably defensible cut scores which took into consideration the students' population. Also the judges involved in the study highly rated and expressed confidence in the procedure utilised.

Recommendations

1. The Borderline Group Method would produce defensible and stable cut scores, even with one-member sub-panels, therefore, the examining bodies (WAEC and NECO) should consider and adopt this design in setting cut scores for grading their candidates.
2. Zonal education authorities in Nigeria should adopt the use of one-member sub-panel BGM standard setting method to set cut scores for assessments of students achievement within their zones

References

Abbott, M. (2003). Standard setting for complex performance assessments: a critical examination of the analytic judgment method. *Running head: The Analytic Judgment Method,* 1, 1-18.

Adewale, J. G. & Antia, I. A. (2016). Standard setting for achievement testing in Nigeria: Which method is practicable? *African Journal of Theory and Practice Of Educational Assessment (AJTPEA)*, 3(1), 66 – 82.

Angoff, W. H. (1971). Scales, norms, and equivalent scores in R. L. Thorndike Eds *Educational measurement*. Washington, DC: American Council on Education. 508–600.

Barman, A. (2008). Standard setting in student assessment: Is a defensible method yet to come? *Annals Academy of Medicine Singapore*, 37(11), 957-63.

Berk, R. A. (1995). Standard setting - The next generation. Joint Conference on Standard Setting for Large-Scale Assessments. Washington, DC: National Assessment Governing Board, National Center for Education Statistics. 2: 161-181.

Cizek, G. J. (2012). Setting performance standards: *Foundations, methods, and innovations.* 2nd ed. New York, NY: Rutledge.

Cizek, G. J. & Bunch, M. B. (2007). Standard setting*: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage Publications Inc.

Cohen-Schotanus, J. & Van der Vleuten, C. P. M. (2010). A standard setting method with the best performing students as point of reference: Practical and affordable. *Medical Teacher,* 32(2),154-60.

Cusimano, M. (1996). Standard-setting in medical education. *Acad Med.*, 71, 112–120.

Downing, S. M. Tekian, A. & Yudkowsky, R. (2006). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and learning in medicine*, 18(1), 50–7.

Elfaki, O. & Salih, M. (2015). Comparison of two standard setting methods in a medical students' MCQS exam in internal medicine, *American Journal of Medicine and Medical Sciences*, 5(4),164-167

Giraud, G. Impara, J. C. & Buckendahl, C. (2000). Making the cut in school districts: alternative methods for setting cut scores. *Educational Assessment,* 6(4), 291 – 304

Gotzmann, A. De Champlain, A. Roy, M. Brailovsky, C. Smee, S. & de Vries, I. (2014). Setting a Common Performance Standard for Two Candidate Groups across Different Medical Licensing Examination Forms: A Comparison of Methods, Technical Report. Medical Council of Canada.

Hambleton, R. K. (2001). Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives.* Mahwah, N.J.: Erlbaum, 89-116.

Hambleton, R. K.. & Pitoniak, M. (2006). *Setting performance standards. Educational Measurement*. Eds R. L. Brennan. Westport, CT: Praeger. 433–470.

Hattie, J. A. C. Brown, G. T. L. & Keegan, P. J. (2003). *Assessment tools for teaching and learning manual*: Volume 2. Auckland: University of Auckland.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15-40.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement,* 38(4), 319-342.

Kane, M. (1998). Criterion bias in examinee-centered standard setting: Some thought experiments. *Educational Measurement: Issues and Practice,* 17(1), 23-30.

Kollias, C. (2012). Standard Setting of the Basic Communication Certificate in English (BCCETM) Examination: Setting a Common European Framework of Reference (CEFR). Technical report. Office for Language Assessment and Test Development, Hellenic American University, New Hampshire, USA.

Livingston, A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

McKinley, D. W. Boulet, J. R. & Hambleton, R. K. (2005). A work-centered approach for setting passing scores on performance-based assessments. *Evaluation & the health professions,* 28(3), 349–69. doi:10.1177/0163278705278282

Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (219–252). Lincloln, NE: Buros Institude of Mental Measurements.

Näsström, G. & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. Practical *Assessment Research & Evaluation,* 13(9). Retrieved May, 02, 2016 from: http://pareonline.net/getvn.asp?v=13&n=9

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

Parmar, D. Shah, C. & Parmar, R. (2014). Study of standard setting in constructed response type written examination. *Int J Med Sci Public Health*, 3(9), 1-5.

Raymond, M. & Reid, J. (2001). Who made thee a judge? Selecting and training participants for standard setting. *Setting performance standards* Eds. G. J. Cizek, New Jersey: Erlbaum, 117 – 158.

Reckase, D. (2010). Study of Best Practices for Vertical Scaling and Standard Setting with Recommendations for FCAT 2.0

Schoeman, F. (2015). Standard Setting for Specialist Physician Examinations in South Africa, Ph.D. HPE Thesis. Health Sciences Education. Health Sciences, University of the Free State, Bloemfontein.