# STANDARD SETTING FOR ACHIEVEMENT TESTING IN NIGERIA: PRACTICABLE AND AFFORDABLE METHODS

**I. A. Antia and J. G. Adewale**

Institute of Education, University of Ibadan

Email: gbengaadewale@yahoo.co.uk

## Abstract

*Formal standard setting methods involving judges in decision making are usually costly and not practicable in small scale testing programs. Formal standard setting methods however seem not to be given much research attention in testing programs in Nigeria. This study was on standard setting for a constructed Physics Achievement Test using the conventional Fixed 50% Pass Mark, Borderline Group Method, an Angoff/Borderline Compromise Method, and the Cohen60 method in Oyo state, Nigeria. The survey research, had a sample of 419 Senior Secondary Two (SS2) students, with their Physics teachers purposively sampled from 10 public secondary schools in three local government areas in Oyo state. Five research questions were answered in the study. Four instruments were constructed and used to collect data. The Physics Achievement Test had reliability ($KR_{20}$ = 0.85). The data obtained were analyzed using descriptive statistics, Item Difficulty, and one way ANOVA. Key findings revealed that The Borderline Group Method, Angoff/Borderline Compromise method, and Cohen60 method yielded comparable cut scores; The Fixed (50%) Pass Mark Method was not sensitive to test difficulty, and not credible; The Cohen60 method was considered as the best performing method; The Borderline Group Method and the Angoff/borderline Compromise method produced comparable pass rates; And the Fixed (50%) Pass Mark Method yielded unacceptable pass rates. It was recommended that the Fixed (50% Pass Mark Method being used in school and colleges examinations should be replaced with the Cohen60 method, Borderline Group Method or the Angoff/Borderline Compromise method, as a particular test situation demands.*

## Introduction

What should constitute the minimum acceptable level of competence (standard) to pass an assessment has been an issue of particular interest and concern to educators mostly in high stakes examinations. Standard setting has been discussed in the literature for more than 50 years and many methods of setting standards have been described and proposed (Abbott, 2003; Barman, 2008; Cusimano, 1996; Cizek, 2007; 2012; Downing, Tekian & Yudkowsky 2006). In broad terms, formal standard setting methods and processes 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).

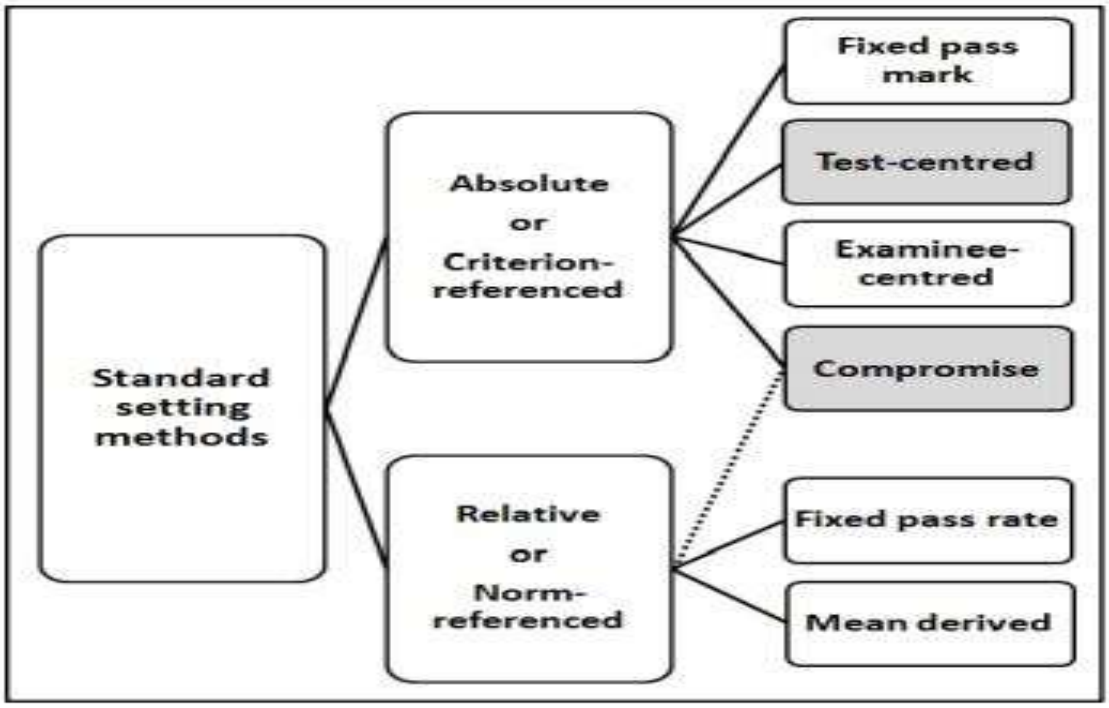
Reckase (2010) gave a more explicit definition of standard setting as the label given to the set of activities that are done to identify points on the reporting score scale for a test that represent desired levels of performance. Standard setting simply put is the process of determining passing scores (cut scores, or performance standards). Further, Hattie and Brown's (2003) explanation that setting performance standards is a process of eliciting reasoned judgments from experts who are (a) knowledgeable about the demands of the test or assessment for which a standard is to be set, (b) understand the meaning of scores at various levels on the scales used to summarize

examinees' performances, and (c) fully comprehend the definitions of achievement associated with the performance standards that they have been asked to establish." Provide more explanation for the practice of standard setting.

Cusimano (1996) asked: "Standard setting is the process of deciding 'what is good enough?' How do we actually make such a decision, when by all conceptions, competence is a continuous variable?" Cusimano (1996) referred to the 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. These conceptions seem to reflect the agreed position of researchers interested in standard setting, with majorly only semantic differences in the different definitions.

The large number of methods for setting performance standards described in the literature (Cizek & Bunch, 2007; Jaeger, 1995; Hambleton & Plake, 1995) can generally be classified into criterion referenced and norm referenced methods, the summary of which is as given by Schoeman, (2015) in Table 1.1

**Figure 1: classification of standard setting methods**




Source: Classification of standard setting methods compiled by Schoeman, (2015)

Norm-referenced standards are considered as the method of choice when the aim is to rank examinees. Criterion-referenced standards are considered most appropriate when the aim is to ascertain whether examinees' mastery of a specific domain meets the pre-set requirements (Norcini 2003). The criterion referenced methods of standard setting in the above classification have been researched and used by various assessors measuring the achievement of learners, for certification, licensure and other purposes. However, as agreed by experts in standard setting, there are no true, objective or "golden" performance standards for any assessment (Kane, 1998),

and the performance standards can only be set in a more or less trustworthy way (Näsström and Nyström, 2008).

In Nigeria, a fixed pass mark of 40% is being used in school assessments at the primary , secondary and tertiary levels of education to classify learners into performance levels pass and fail. In this system, 50% serves as the cut score for a level of pass referred to as the credit pass, which represents minimum level of performance considered for competence in a given assessment. This will be adopted in this study, resulting in two performance levels credit pass and above, and below credit pass. A credit pass in school subject say chemistry at ordinary (O') level for instance implies that the learner has mastered the contents of chemistry creditably and so considered competent in chemistry at that level.

However, the use of the fixed pass mark in such decisions has been severally criticized by researchers, and experts in standard setting. Searle (2000:363) and Bhandary (2011:3), argue strongly that although it is easy to set a pass mark at 50%, it is not a fair measure to determine who is competent or not and that it is not transparent nor is it defensible. The concern with this method is that there is no link to the standard of the test and it is completely insensitive to the difficulty of the test, (Schoeman 2011:2; Van der Vleuten 2010:175), which means it has an unknown relationship with competence (Searle 2000:366). It is not credible, since any standard setting method that does not take test difficulty into account in some way is in danger of damaging credibility (Cohen-Schotanus & Van der Vleuten, 2010).

To further demonstrate the insensitivity of this method to the difficulty of the test, if a student sits for a test, which contains 100 test items, and scores 60/100 (60%) on the test. Using the 50% fixed mark as cut score for competence, will classify the student as competent, since the student obtained a score of more than 50%. However, this judgment might seem inappropriate if 200 students wrote the same test and the class average was 80%. This result would imply that the test was easy for the group of students and hence a score of 60% should have been classified differently. This method is mostly based on tradition (Zieky 1995:33).

This advocates for the other two criterion referenced methods; test centered, and examinee centered. These also are not unalloyed blessings. The Criterion-referenced standard setting procedures typically require panels to determine the minimally acceptable level per item (Bandaranayake 2008). However, panel procedures are time consuming and, therefore, often too costly to use for in-house tests. The generally limited resources prohibit the regular use of panels for standard setting procedures (Cohen-Schotanus & Van der Vleuten, 2010). This explains the adherence to the crude fixed pass mark method used in small scale testings.

In order to minimise the disadvantages of both the norm referenced and the criterion referenced methods, compromise methods were proposed. A compromise method, combining a pre-fixed (criterion referenced) cut-off score with a relative (norm referenced) 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).

This study sought to find more footing for the adoption of credible standard setting methods over the use of the crude fixed pass mark method used by assessors in Nigeria. The study empirically compared the performance of four methods of standard setting; the conventional 50% fixed pass

mark method, a version of Borderline Group Method, an Angoff/Borderline Compromise Method, and the Cohen method using a constructed Physics Achievement Test (PAT) in Oyo state, Nigeria.

**The Conventional 50% Fixed Pass Mark Method:** this is the conventional method, in which the cut score for the PAT is preset at 50% of the total obtainable score in the test to classify students into two performance levels credit pass and above and below credit pass. The total obtainable score for the PAT is 70 marks hence the 50% fixed pass mark cut score is preset at 35. Therefore, students scoring 35 and above are graded with credit pass and above, while students scoring below 35 are graded with below credit pass.

**The Borderline Group Method:** The Borderline Group Method is based on the idea that the cut score should be the score that would be expected from a test taker whose skills are "on the borderline" — not quite adequate and yet not really inadequate. This method calls for the judges to identify actual test takers as "borderline" in the knowledge and skills the test measures. The judges do not have to judge all of the test takers or even a representative sample of them. They need only identify the ones who, in their judgment, best fit the definition of a borderline test taker. The cut score is then set at the median score (the 50th percentile) of this "borderline group" (Livingston & Zieky, 1982).

This method usually involves a group of seven to ten or more judges, in a workshop organised for some two three or more days in order to arrive at a cut score for a given assessment. However in this study, the training as well as the judgment processes were decentralised, and the judges selected for the study were trained separately, and they made their judgments separately. Beyond the practicability and cost issues, this modification is believed to yield good results, in line with Raymond and Reid's, (2001) main criteria for judge selection; the judges' familiarity with (1) the examinee population; and (2) the intended performance levels to be set. Since the judges who are teachers will be very familiar with the test takers their students (only teachers who have been with the students for a period of at least one academic session were selected). The teachers were required to identify borderline test takers among their students presented for the test. And the cut score is to be set at the median test score of the identified borderline group students.

**The Angoff/Borderline Compromise Method:** As the name suggests, this is a method conceived from the ideas of both the Angoff/modified Angoff and the Borderline Group Methods. This method involves the determination through judgment by the judges of the borderline group test takers, administration of the test to the sampled examinees, and setting the cut score at the total item difficulty (test difficulty) using data from the borderline group examinees. In so doing, the method borrows from the two parent methods; the Borderline Group Method and the Angoff Method. The use of the median score of the borderline group examinees in the Borderline Group Method is replaced by the use of the total item difficulty of the borderline examinees. The total item difficulty used is arrived at from the judgment of judges in the Modified Angoff Method; in which judges are asked to "picture 100 borderline students and determine how many of them would answer the item correctly" This concept of judgment is replaced by the actual test difficulty of the test, making use of the borderline examinees' test scores. This method was put together by Adewale and Antia, 2016

**The Cohen Method:** The Cohen method of standard settingwas first published in the Dutch literature in 1996 (Cohen-Schotanus, Van der Vleuten & Bender 1996:83 -87), and

then in the English literature in 2010 (Cohen-Schotanus & Van der Vleuten 2010:157 ), holds much promise as a cost-effective and sustainable tool to determine the pass mark of summative examinations in a resource-limited setting (Schoeman, 2015).

In the Cohen method, the top-performing students are used as a point of reference to set an absolute pass mark. Essentially, the performance of the top candidates (90 - 95$^{th}$ percentile of the test scores) is used as the benchmark for the difficulty of the assessment and the pass mark is usually set at 60-70% of the benchmark (Cohen-Schotanus & Van der Vleuten 2010:159 ). The 95$^{th}$ percentile is usually used because available research data suggests that this top cohort of candidates is quite stable and performs equally well between different cohorts of examinees as compared to the mean test score, which is dragged down by poorly performing students (Cohen-Schotanus & Van der Vleuten 2010:157). In addition, the use of the 95$^{th}$ percentile also makes provision for 5% top-end outlier performers and hence, the outliers do not affect the pass standard for the cohort under assessment. In this study, the cut score will be set at 60% of the benchmark (95$^{th}$ percentile). This will be called Cohen60.

**Statement of the Problem**

Standard setting has been researched and discussed in literature for over 50 years. Many methods have been compared for different forms and purposes of examinations. Standard setting methods involving the use of judges have cost and practicability implications on small scale testing programs.

In Nigeria, crude methods of standard setting are still being traditionally held on to. This study therefore sought to find more footing for the adoption of more credible standard setting methods over the use of the crude fixed pass mark method being used by assessors in Nigeria.

**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 cut scores set for the PAT using the four methods for the different schools?
3. What pass rates obtained in the different schools as a result of the cut scores set by the four methods?
4. Is there any significant difference in the cut scores set by the four methods for the PAT in the different schools?
5. Is there any significant difference in the pass rate obtained in the different schools as a result of the cut scores set by the four methods for the PAT?

**Aim of the Study**

The aim of this study is to empirically compare the performance of four standard setting methods using a constructed Physics Achievement Test in Oyo State, Nigeria.

More specifically,

- Evaluate the cut scores set for the PAT by the four methods for the different groups of test takers (schools)

- Compare the consequences (pass rates) of the resulting cuts cores of the four methods on the group of test takers.

## Methodology

### Research Type

The present study was a survey, involving Standard setting methods comparison.

### Population

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.

### Sampling Techniques and Sample

The researcher made use of purposive sampling technique for selecting the schools that participated in this study. Since the study involved a lot of communication between the researcher and the teachers (judges), there was the 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 in Oyo State that were close together (Ibadan north, Ibadan northwest, and Ibadan north east) were purposively selected. Ten schools that were as close as possible, after leaving out those that did not meet the criteria were also purposively selected. All the available SS2 students in the selected schools participated in the study. The total sample size was 419 students. The Physics teachers of the selected students in the 10 schools served as judges in the study.

### Instrumentation

The instruments that were used in gathering data for this study included;

### Physics Achievement Test (PAT)

The PAT was divided into two parts (examinations), similar to the objective and essay parts in the WASSCE Physics. The paper I consisted of 1 – 50 multiple choice questions, with four options lettered A – D and paper II consisted of 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.

**Table 1: Table of Specification for Physics Achievement Test**

| | Knowledge | Comprehension | Higher order | Total (%) |
|---|---|---|---|---|
| theme1 | 2 | 3 | 5 | 10 (18) |
| theme2 | 6 | 7 | 14 | 27 (49) |
| theme3 | 4 | 7 | 5 | 16 (29) |
| theme4 | - | - | - | - |
| theme5 | 1 | - | 1 | 2 (4) |
| theme6 | - | - | - | - |
| Total (%) | 13 (24) | 17 (31) | 25 (45) | 55 (100) |

**Training Manual for the Borderline Group Standard Setting Method (TMBGSSM)**

This consisted of introduction to standard setting, information about the students' population, and test, detailed explanation of the Borderline Group Method, and a familiarization task for participant judges to practice the actual exercise with.

**Borderline Group Sheet (BGS)**

This instrument was also 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 his judgment.

**Borderline Group Method Evaluation Sheet (BGMES)**

This 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.

**Validation of Instrument**

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.

**Data Collection Procedure**

The researcher first visited the sampled schools, and teachers, seeking their consent and permission for conducting the study. The selected judges were trained in the use of the Borderline Group Method through the training manual (individually). The judges were then allowed to go through and complete the familiarization task, and their understanding of the process was ensured. The judges were then required to complete the Borderline Group Method Evaluation Sheet (BGMES), rating their confidence in the process. The students were then informed of the test date and asked to prepare. 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. The students were allowed enough time to complete the test, and then the materials were retrieved for marking and further data analyses. The judges were later given feedback on their students' achievement on the test. Three research assistants were trained and used as invigilators during the test administrations. The data collection lasted for three weeks

**Data Analyses:**

The data collected were analyzed as shown below:

**Table 2: Data Analyses**

| S/N | Research Questions | Statistical Analysis |
|---|---|---|
| 1 | 1 | Minimum score, maximum score, mean score, and standard deviation. |
| 2 | 2, and 3 | Percentages, percentiles, Median and item difficulty. |
| 3 | 4, and 5 | One way ANOVA |

## Results

### Research Question 1

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

**Table 3: Raw Scores Statistics for the Physics Achievement Test (PAT)**

| School | N | Min. Score | Max. Score[a] | Mean score | Standard deviation |
|---|---|---|---|---|---|
| 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 | 34 | 9 | 29 | 16 | 4.21 |
| school 6 | 43 | 10 | 27 | 18 | 4.30 |
| school 7 | 23 | 8 | 25 | 16 | 4.31 |
| school 8 | 28 | 14 | 48 | 33 | 8.77 |
| school 9 | 45 | 7 | 28 | 19 | 4.94 |
| school 10 | 40 | 17 | 41 | 32 | 5.14 |

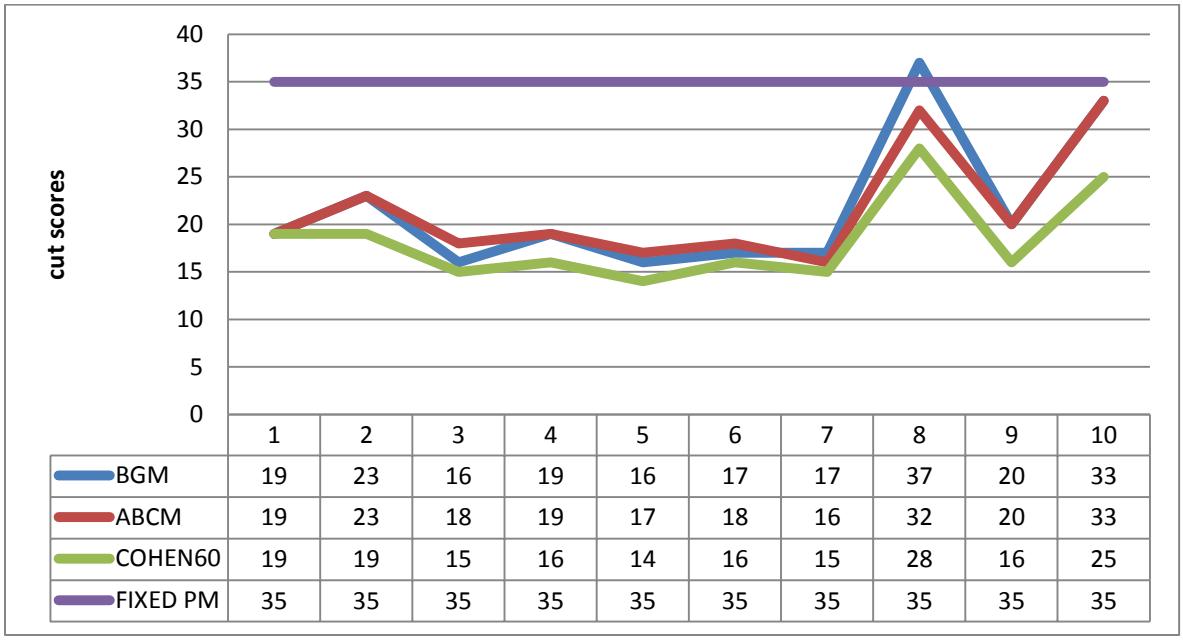
N = number of students; a = the maximum obtainable score on the test was 70

Result from Table 4.1 above shows the 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. In schools 3, 5, and 7, the lowest mean test scores of 16 were obtained, with standard deviations of 4.24, 4.21, and 4.31, minimum scores of 9, 9, and 8. And maximum scores of 25, 29, and 25 respectively. On the other hand, schools 8, and 10 had the highest class mean test scores of 33, and 32, with standard deviations of 8.77, and 5.14, minimum scores of 14 and 17. And maximum scores of 48 and 41 respectively.

### Research Question 2

What are the cut scores set for the PAT using the four methods for the different schools?

**Figure 2: Cut scores set for the PAT using the four methods using the four methods for the different schools**




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| BGM | 19 | 23 | 16 | 19 | 16 | 17 | 17 | 37 | 20 | 33 |
| ABCM | 19 | 23 | 18 | 19 | 17 | 18 | 16 | 32 | 20 | 33 |
| COHEN60 | 19 | 19 | 15 | 16 | 14 | 16 | 15 | 28 | 16 | 25 |
| FIXED PM | 35 | 35 | 35 | 35 | 35 | 35 | 35 | 35 | 35 | 35 |

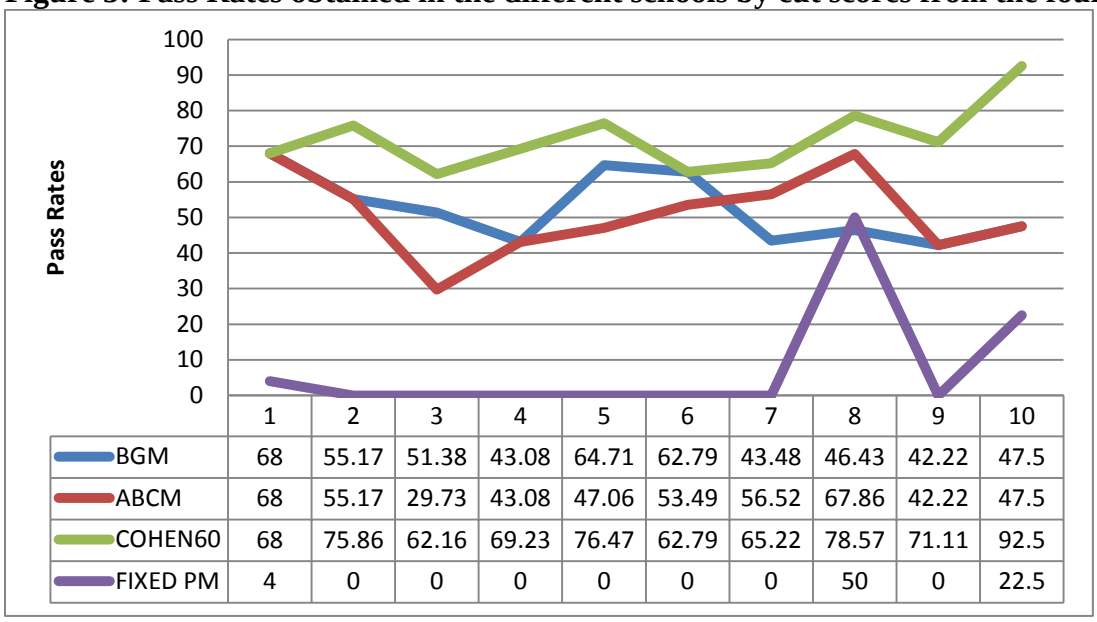
Key: BGM: Borderline Group Method; ABCM: Angoff/Borderline Compromise method; COHEN60: Cohen60 method; FIXED PM: Fixed Pass Mark (50%) method

Result in Figure 2 shows the cut scores set for the PAT using the four methods for the different groups of test takers. The Fixed Pass Mark (50%) method gave a cut score of 35 across the ten schools taking the test. the Borderline Group Method, Angoff/Borderline Compromise method, and Cohen60 method gave cut scores having fluctuations, which are somewhat regular. The Cohen60 method consistently gave lower cut scores than the Borderline Group Method and the Angoff/Borderline Compromise Methods, except for school 1 where the three methods gave the same cut score of 19. The Borderline Group Method and the Angoff/Borderline Compromise Method cut score were the same for schools 1, 2, 4, 9, and 10, with the others fluctuating with little differences between the two methods

**Research Question 3**

What pass rates obtained in the different schools as a result of the cut scores set by the four methods?

**Figure 3: Pass Rates obtained in the different schools by cut scores from the four methods**




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| BGM | 68 | 55.17 | 51.38 | 43.08 | 64.71 | 62.79 | 43.48 | 46.43 | 42.22 | 47.5 |
| ABCM | 68 | 55.17 | 29.73 | 43.08 | 47.06 | 53.49 | 56.52 | 67.86 | 42.22 | 47.5 |
| COHEN60 | 68 | 75.86 | 62.16 | 69.23 | 76.47 | 62.79 | 65.22 | 78.57 | 71.11 | 92.5 |
| FIXED PM | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 22.5 |

Result in Figure 3 shows pass rates obtained as a result of the cut scores set by the four methods for the different groups of test takers. Pass rates of 4%, 50%, and 22.5% resulted from cut scores set by the Fixed Pass Mark (50%) for schools 1, 8, and 10 respectively. The Borderline Group Method, Angoff/Borderline Compromise method, and Cohen60 method gave cut scores resulting in pass rates with fluctuations. The Cohen60 cut scores resulted in higher pass rates than the Borderline Group Method and the Angoff/Borderline Compromise Methods, except for school 1 where the three methods gave the same cut score, and therefore pass rates. The Borderline Group Method and the Angoff/Borderline Compromise Method cut score were the same for schools 1, 2, 4, 9, and 10, and therefore the pass rates were the same, with the others fluctuating with little differences between the two methods.

**Research Question 4**

Is there any significant difference in the cut scores set by the four methods for the PAT in the different schools?

**Table 4: Difference in the cut scores set by the four methods for the PAT**

**ANOVA**

DV: CUT SCORES

| | Sum of Squares | Df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 1649.675 | 3 | 549.89 | 19.39 | .00 |
| Within Groups | 1020.700 | 36 | 28.35 | | |
| Total | 2670.375 | 39 | | | |

Table 4 reveals the difference in the cut scores set by the four methods for the PAT. From the table, the F-value, 19.395 is significant at 0.05, ($P<0.05$). It follows that there is significant difference in the cut scores set for the PAT by the four methods.

**Table 5: Post-hoc test on the difference in cut scores set by the four methods for the PAT**

**Multiple Comparisons**

Dependent Variable: CUT SCORES

Tukey HSD

| (I) METHOD | | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval Lower Bound | 95% Confidence Interval Upper Bound |
|---|---|---|---|---|---|---|
| BGM | ABGM | .20 | 2.38 | 1.00 | -6.21 | 6.61 |
| | COHEN60 | 3.40 | 2.38 | .49 | -3.01 | 9.81 |
| | FIXED PM | -13.30* | 2.38 | .00 | -19.71 | -6.89 |
| ABGM | BGM | -.20 | 2.38 | 1.00 | -6.61 | 6.21 |
| | COHEN60 | 3.20 | 2.38 | .54 | -3.21 | 9.61 |
| | FIXED PM | -13.50* | 2.38 | .00 | -19.91 | -7.09 |
| COHEN60 | BGM | -3.40 | 2.38 | .49 | -9.81 | 3.01 |
| | ABGM | -3.20 | 2.38 | .54 | -9.61 | 3.21 |
| | FIXED PM | -16.70* | 2.38 | .00 | -23.11 | -10.29 |
| FIXED PM | BGM | 13.30* | 2.38 | .00 | 6.89 | 19.71 |
| | ABGM | 13.50* | 2.38 | .00 | 7.09 | 19.91 |
| | COHEN60 | 16.70* | 2.38 | .00 | 10.29 | 23.11 |

*. The mean difference is significant at the 0.05 level.

The post-hoc table shows that the significant difference in cut scores among the four methods ($p<0.05$) is as a result of the high mean difference between the cut scores set by the Fixed (50%) Pass Mark Method and the other 3 methods (13.30, 13.50, and 16.70 between the BGM, ABCM, and COHEN60 respectively), which were the only significant differences.

**Research Question 5**

Is there any significant difference in the pass rate obtained in the different schools as a result of the cut scores set by the four methods for the PAT?

**Table 6: Difference in Pass Rates obtained as a result of cut scores set by the four methods for the PAT**

**ANOVA**

DV: PASS RATES

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Between Groups | 22241.674 | 3 | 7413.89 | 50.59 | .00 |
| Within Groups | 5275.535 | 36 | 146.54 | | |
| Total | 27517.209 | 39 | | | |

Result in Table 6 shows the difference in the pass rate obtained as a result of the cut scores set by the four methods for the PAT. From the table, the F-value, 50.592 is significant at 0.05, ($p<0.05$). It follows that there is significant difference in the pass rate obtained as a result of the cut scores set by the four methods for the PAT.

**Table 7: Post-hoc test on the difference in pass rate obtained as a result of the cut scores set by the four methods for the PAT**

**Multiple Comparisons**

Dependent Variable: PASS RATES

Tukey HSD

| (I) METHOD | | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval Lower Bound | 95% Confidence Interval Upper Bound |
|---|---|---|---|---|---|---|
| BGM | ABGM | 1.41 | 5.41 | .99 | -13.17 | 15.99 |
| | COHEN60 | -19.72* | 5.41 | .00 | -34.30 | -5.13 |
| | FIXED PM | 44.83* | 5.41 | .00 | 30.25 | 59.41 |
| ABGM | BGM | -1.41 | 5.41 | .99 | -15.99 | 13.17 |
| | COHEN60 | -21.13* | 5.41 | .00 | -35.71 | -6.55 |
| | FIXED PM | 43.41* | 5.41 | .00 | 28.83 | 57.99 |
| COHEN60 | BGM | 19.72* | 5.41 | .00 | 5.13 | 34.30 |
| | ABGM | 21.13* | 5.41 | .00 | 6.55 | 35.71 |
| | FIXED PM | 64.54* | 5.41 | .00 | 49.96 | 79.12 |
| FIXED PM | BGM | -44.83* | 5.41 | .00 | -59.40 | -30.25 |
| | ABGM | -43.41* | 5.41 | .00 | -57.99 | -28.83 |
| | COHEN60 | -64.54* | 5.41 | .00 | -79.12 | -49.96 |

*. The mean difference is significant at the 0.05 level.

The post-hoc table shows that the significant difference in Pass Rates obtained among the four methods ($p<0.05$) is as a result of the high mean difference in Pass Rates obtained between the Fixed (50%) Pass Mark Method and the other 3 methods (-44.83, -43.41, and -64.54, between

the BGM, ABCM, and COHEN60 respectively) Also, the high mean differences in Pass Rates between the Cohen 60 Method and the other two methods (19.72, and 21.13, between the BGM, and ABCM respectively). These were the only significant differences.

**Discussion**

Research question one was aimed at providing statistical information on the performance of the students taking the test. It was seen from the variation of the mean scores that the difficulty of the test varied among the sampled schools. As presented in Table 3, schools 3, 5, and 7 obtained the lowest mean scores. This means that the test was most difficult for the students in these schools. This could be as a result of their teachers not teaching them to mastery in the contents from which the test items were drawn, or other similar factors. Their low and comparable standard deviations imply that the students taking the test in these three schools obtained scores which bunch around their low mean scores with few outlying scores. This is also evident in the little differences between the minimum scores and the mean scores and also between the maximum scores and the mean scores. This will bring down the cut scores set for these schools by the methods that are test difficulty sensitive. These schools had students whose achievements in the test were more homogeneous.

On the other hand, schools 8, and 10 highest mean scores indicate that the test was less difficult or easiest for the students in these schools. This could be attributed to the teaching and learning variables/situation in the two schools. School 8 recorded the highest standard deviation, which suggests that the students' scores had the widest spread about the mean score in this school, hence the wide difference between the minimum score and the mean score as well as between the maximum score and the mean score. This shows less homogeneous achievement and the presence of greatly outlying scorers. Higher cut scores are expected for schools 8, and 10 with the use of the test difficulty sensitive methods.

Result presented in figure 2 shows a cut score of 35 set using the Fixed (50%) Pass Mark Method for all the sampled schools, and fluctuating cut scores set by the Borderline Group Method, Angoff/Borderline Compromise method, and Cohen60 method for the different schools. Comparing these fluctuations with result in table 3 reveal that lower cut scores are set for schools for which the test was more difficult (schools 3, 5, 7), and higher cut scores were set for schools for which the test was less difficult (schools 8, and 10). This means that these methods are sensitive to the test difficulty for a particular cohort (school), while the Fixed (50%) Pass Mark Method is rigid and insensitive to test difficulty. This agrees with the concerns of Schoeman, (2011) and Van der Vleuten, (2010), hence the Fixed (50%) Pass Mark Method continues to prove not credible Cohen-Schotanus & Van der Vleuten, (2010).

Result in Table 4 shows significant difference in the cut scores set by the four methods for the PAT. Also, multiple comparisons in table 5 reveals that the difference is caused by the high cut scores set by the Fixed (50%) Pass Mark Method, which results in high mean differences between the Fixed (50%) Pass Mark Method and the other three methods. The mean differences of the other three methods not being significant implies that cut scores set by these methods are comparable, although the Cohen60 method produces the lowest cut scores among the three.

The consequences of the various cut scores set by the four methods for the various schools, as presented in figure 3 show that the Fixed (50%) Pass Mark Method the pass rates in 7 out of the

10 sampled schools (schools 2, 3, 4, 5, 6, 7, 9) to be 0%. This is unacceptable seeing that this did not consider the difficulty of the test in these schools. The fluctuations observed in the pass rates obtained by the other three methods are also traceable to their considerations for the test difficulties in the various schools as in the case of the underlying cut scores discussed above.

However, result in table 6 showing significant difference in the pass rates obtained by the four methods' cut scores is interesting. As seen in the multiple comparisons of table 7, the difference is as a result both of the unacceptably low pass rates obtained from the Fixed (50%) Pass Mark Method cut score, and the high pass rates obtained from the Cohen60 method. The Cohen60 method though yielding comparable cut scores with the Borderline Group Method and the Angoff/Borderline Compromise Methods produces a significantly higher pass rates than the two of them.

## Conclusion

The main inferences drawn from this study were that the difficulty of the test varied among the sampled schools. The Borderline Group Method, Angoff/Borderline Compromise method, and Cohen60 method yielded comparable cut scores, with the Cohen 60 method producing the lowest cut scores consistently. The Fixed (50%) Pass Mark Method is not sensitive to test difficulty, and so not credible. Also, it was gathered that the Cohen60 method yielded higher pass rated compared with the other difficulty sensitive methods, and so can be considered as the best performing method, coupled with its cost effective and practicable nature. The Borderline Group Method and the Angoff/borderline Compromise method also produced acceptable and comparable pass rates. And finally the Fixed (50%) Pass Mark Method yielded unacceptable pass rates.

## Recommendations

Based on the findings of the study, the following recommendations were made

- The Cohen 60 method should be adopted for use in the classification of test takers achievement in small scale testing programs such as school examinations, board examinations, professional certifying or licensure examination in Nigeria.
- The Borderline Group Method or Angoff/Borderline Compromise method should be used in classification of test takers achievement in both large scale and small scale testing programs if it is possible for the tester to reliably judge the test takers into performance levels.
- The Fixed (50%) Pass Mark Method in use in school and colleges examinations should be replaced with the Cohen 60 method or either of the Borderline Group Method or the Angoff/Borderline Compromise method, as applicable in situation.

## References


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

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

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.* 1996; 71:112–120.

Hambleton, R. K., & Plake, B. S. 1995.Using an extended Angoff procedure to set standards on complex performance assessments. *Appled Measurement in Education*, 8, 45–55.

Jaeger, R. M. 1989.Certification of student competence. In R. L. Linn (Ed.) *Educational Measurement* 3rd ed. New York: American Council of Education & McMillan.

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

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.

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

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.