AFRICAN JOURNAL OF THEORY AND PRACTICE OF EDUCATIONAL ASSESSMENT (AJTPEA)

VOL 8(1), 2019: 124 – 139 Copyright © Educational Assessment and Research Network in Africa (EARNiA). ISSN 2545-5834
http://www.earnia.org; info@earnia.org; earnia13@gmail.com

A COMPARISON OF JUGEMENTAL AND NON-JUDGEMENTAL METHODS OF STANDARD SETTING FOR PHYSICS ACHEIVEMENT TEST USING ONE-MEMBER SUB PANELS IN SECONDARY SCHOOLS IN OYO STATE, NIGERIA

Inimfon A. Antia, & Juliana F. Udoudoh

Abstract

Standard setting is the process of determining passing scores, or cut scores, or performance standards for a given psychological measure. Developers and users of psychological measures in Nigeria seem not to bother about formal standard setting practices. This study was designed to compare a non judgemental method of standard setting (Cohen method) with two judgemental methods (Borderline group method, and an Angoff/Borderline compromise method) working with one-member sub panels in Secondary schools in *Oyo state, Nigeria. The survey research, from a population of all secondary schools in Oyo sate, sampled 438 Senior Secondary Two (SS2) students, with their Physics teachers purposively sampled from 11public secondary schools in three local government areas of Oyo state. Four instruments were constructed, validated and used for data collection. Data were analyzed using descriptive statistics (*minimum scores, maximum scores, mean scores and standard deviation*), Item Difficulty, and one way ANOVA. Findings revealed that the three methods produced reasonable and useful cut scores, although the Borderline Group Method produced unstable individual cut scores with smaller sample sizes. On the grounds of the findings, it was recommended that the Cohen 60 method, Angoff/Borderline Group Method or Borderline Group Method should be adopted by testing organizations and teachers in setting cut scores for grading their candidates'/students' test.*

Introduction

Tests and other psychological measures are widely used in the fields of education and social sciences to quantify several characteristics of individuals, items or events. A test for instance can be designed to measure the achievement of students in a school subject, after the students might have been exposed to some learning opportunity. The test so designed attempts to do what the measuring tape for instance does in the measurement of physical dimensions of length. The tape provides information using an internationally recognized unit say “centimeters” on the dimension such as height of a maize plant, for example, the measurement of a particular plant could be given as 75 centimeters (cm). This single measurement or even several random measurements may not convey much information about the maize plant since there is no expectation.

Now let us assume that from standard maize farming practice, the maize farmer expects four weeks old maize plants to have heights ranging between 100cm and 120cm, with this expectation serving as a standard, measurements of the height of four weeks old maize plant would tell him the situation of the maize plants in his farm, and will direct him on which plant to think of probably applying a growth boosting fertilizer, or which not to apply, supposing that the height of the plants was the major indicator of their wellbeing. Two boundaries of the plants' height are therefore set (at 100cm and 120cm), which could be used to categorize the maize plants into three categories short (below 100cm), moderate (100cm to 120cm) and tall (above 120cm).

Similar to the physical measure illustrated above with the maize plants' height measurement, psychological measurements make little sense to the stakeholders without standards. A questionnaire measuring self-motivation of certain category of learners for instance should be able at the end of the measurement to say; individuals who respond to the instrument at such and such levels should be classifiable into such and such categories of self-motivation. To achieve this kind of classification, expectations or standards (as is the case with the physical measurement example) must be established in the process referred to as standard setting.

Standard setting has been a topical issue in measurement and assessment over the past 50 years. Standard setting in the most simplified terms is the process of determining passing scores (cut scores, or performance standards). Reckase, (2010) defines standard setting as the name ascribed to the set of outlined procedures carried out to identify locations or points on the reporting score scale for a test that indicate several desired levels of achievement or performance. Cusimano, (1996) described the word “standard” as a conceptual boundary (on the true-score scale) between adequate performance and inadequate performance, and a passing score (cut score) on the other hand as a definite point (on an observed-score scale) that serves as a reference point for decisions making about examinees performance. These conceptions seem to reflect the agreed position of researchers interested in standard setting, with majorly only semantic differences in the various definitions.

Several approaches to, or methods of setting standards have been described and proposed by previous research (Cizek, 2012; Barman, 2008; Downing, Tekian & Yudkowsky 2006; Abbott, 2003; Cusimano, 1996). 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). The escalating number of standard setting methods of standard setting in previous studies (Cizek & Bunch, 2007; Jaeger, 1995; Hambleton & Plake, 1995) can generally be dichotomized into criterion referenced methods and norm referenced methods (Schoeman, 2015). Some of the commonly used methods of standard setting include: the Angoff method, Nedelsky method, Ebel method, borderline group method, Cohen method, and Hofstee method among others. In the present study, the borderline group method, Cohen method and the Angoff method, which make up most part of the standard setting literature, will be considered.

The Borderline Group Method (BGM): According to Livingston & Zieky, in Adewale and Antia, (2016) the Borderline Group Method stipulates that the cut score is the score expected from a test taker whose skills are said to be "on the borderline". To execute the BGM, a group of seven to ten or more judges are usually required to participate in a three-day or more workshop organized to arrive at a cut score for a given assessment. The present study was interested in decentralizing the standard setting process, hence the individual judges will be treated as one-member sub-panels, therefore the training as well as the judgement processes were decentralized, and the judges selected for the study were trained separately, and they made their judgements separately. This adjustment was occasioned by practicability considerations, however, the modification was consistent with Raymond and Reid's, (2001) main criteria for judge selection, since the judges who are teachers will be very familiar with the test takers (their students). For more detailed description of the BGM, see Livingston and Zieky, (1982).

The Angoff/Borderline Compromise Method (ABCM): This method was conceived from the ideas of both the Angoff/modified Angoff and the Borderline Group Methods by Adewale and Antia, (2016). It involves the determination 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. The method consequently borrows from the two parent methods; the BGM and the Angoff Method. In the ABCM, the median score used in the BGM 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 adjudged borderline examinees' test scores (Adewale & Antia, 2016).

The Cohen Method: This method first appeared in the Dutch literature in 1996 (Cohen-Schotanus, Van der Vleuten & Bender 1996), and then in the English literature in 2010 (Cohen-Schotanus & Van der Vleuten 2010).The method is seen as a cost-effective and sustainable tool to determine the pass mark of summative examinations in a resource-limited setting (Schoeman, 2015). With the Cohen method, the performance of the top candidates (90 - 95th 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). The 95th 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 with the mean test score, which is dragged down by poorly performing students (Cohen-Schotanus & Van der Vleuten 2010). The present study, was carried out with the cut score set at 60% of the benchmark (95th percentile), and this was called Cohen60 method.

Standard setting methods which make use of judges always come with high cost and practicability issues, especially in small scale testing programmes (Adewale & Antia, 2016). And almost all the criterion-referenced methods of standard setting involve judges. The constitution of a large, representative, qualified and credible panel of expert judges is an important component of a standard setting study (Bahry, Hambleton, Gotzmann, De Champlain, & Roy, 2012) because of the validity and generalizability of the decisions to be made (Berk, 1995). Qualification in this context is determined on a number of criteria. Raymond & Reid, (2001) used as main criteria for judge selection, the judges' familiarity with (1) the examinee population; and (2) the intended performance levels to be set. This can be directly measured by the number of years of experience of the judges in; working with the examinee population, measurement and assessment, working with examination boards. All these are done in a bit to ensure the validity of the process and whatever standards emerge from the process.

The other important consideration in selection of judges is the number or size of the judges' team or panelist. Several recommendations are made by researchers, such as 15 to 20 persons (Jaeger, 1991; Hambleton, Jaeger, Plake, and Mills, (2000). However, some studies have reported the use of panelists less than 15 members though the minimum often do not fall below 3-6 judges (Norcini et al., 1993). To ensure and demonstrate the validity and generalizability of performance standards set, Hambleton, (2001) outlined three designs of constituting the panel of judges. The first design involves doubling the size of appropriate number of judges for the given situation, the second involves constituting two separate panels of judges of about equal sizes, while the third stipulates creation of subpanels from a single appropriate sized sample of judges. Generalizability of performance standards over judges' panel, which is the aim of implementing any these designs, seeks to ensure that if another group of panelists with the same characteristics as the first were to be constituted, that the set of performance standards that would be obtained would be similar to those of the first panel.

The third design, which the present study adopted and other similar developments in standards setting are aimed at reducing the cost of achieving credible standards setting while upholding important generalizability as well as validity requirements, rather than give in to cheap yet indefensible traditional methods. In the present study, the subpanels design was further modified to one-member subpanels. The aim of the study was to empirically compare the performance of three standard setting methods using a constructed Physics Achievement Test in secondary schools in Oyo State, Nigeria. In line with Raymond & Reid's, (2001) conditions for competence, and credibility (qualification) the profiles of the judges that participated in the study was examined. Also, the validity and generalizability of the cut scores that were obtained from the overall panel were tested using the agreement of the outcome (percentage pass) of sub-panels cut-scores with that of the overall panel. Two judgemental methods of standard setting, The BGM, and ABCM were compared with a third method (Cohen 60) which does not require panel judgment (non-judgemental), in terms of their cut scores, percentage pass, and generalizability of cut scores over the judges panel, working with one-member sub panels for classifying students' achievements in the PAT.

Research Questions

1. What are the raw score statistics (minimum scores, maximum scores, mean scores and standard deviation) for the Physics Achievement Test (PAT)?
2. What are the profiles of the judges involved in the standard setting processes?
3. What are the outputs (cut scores) and consequences (percentage pass) of the three methods of standard setting for the PAT in the different schools and for the overall judges' panel?
4. Are there significant differences among the performance of the three methods in terms of
   a. Cut Scores?
   b. Percentage Pass?
5. To what extent are the cut-scores set by the three methods generalizable over the judges' panel used?

Methods

The present study was a survey, involving Standard setting methods comparison. 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 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 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, 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 438 students. The Physics teachers of the selected students in the 11 schools served as judges in the study.

The instruments that were used in gathering data for this study included the Physics Achievement Test (PAT), which 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 items, with four options lettered A – D and paper II consisted essay items having five (5) short answered questions to answer all for 20 marks, making the total test maximum obtainable score to be 70 marks. Other instruments included the Training Manual for the Borderline Group Standard Setting Method (TMBGSSM), Borderline Group Sheet (BGS) and the Borderline Group Method Evaluation Sheet (BGMES). 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 review.

To collect data for the study, the researcher first visited the sampled schools, and teachers, seeking their consent and permission to conduct the study. The selected teachers who served as 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. During the administration of the test, 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 their scripts were retrieved for marking and further data analyses. The data collection lasted for three weeks. The data collected were analyzed using Minimum score, maximum score, mean score, and standard deviation, Percentages, percentiles, Median and item difficulty, One way ANOVA and paired sample t-test.

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 1: 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	19	12	30	21	4.28
school 11	40	17	41	32	5.14
All schools	438	7	48	20	7.44

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

Result from Table 1 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 11 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. The table also revealed that the overall minimum score for the test was 7, while the overall maximum score was 48. The overall test sample also had a mean score of 20, with a standard deviation of 7.44. The results presented in the table demonstrated that the raw scores obtained from the administration of the PAT to students from the different schools, varied widely across the schools.

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

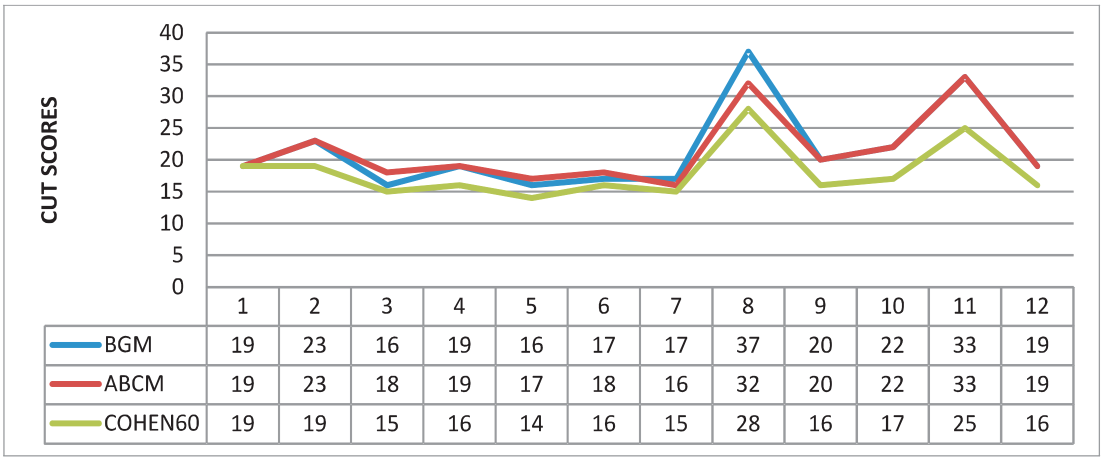
Table 2: Profiles of Judges Involved in the Standard Setting Exercise

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
J3	H.N.D, B. Sc, B, Ed/ equivalent	Below 5	Below 5	1 year and above
J4	H.N.D, B. Sc, B, Ed/ equivalent	5 and Above	5 and Above	1 year and above
J5	H.N.D, B. Sc, B, Ed/ equivalent	5 and Above	Below 5	1 year and above
J6	POSTGRADUATE DEGREE	Below 5	Below 5	below 1
J7	H. ND, B. Sc, B, Ed/ equivalent	5 and Above	5 and Above	1 year and above
J8	POSTGRADUATE DEGREE	5 and Above	5 and Above	1 year and above
J9	POSTGRADUATE DEGREE	5 and Above	5 and Above	1 year and above
J10	H. ND, B. Sc, B, Ed/ equivalent	5 and Above	5 and Above	1 year and above
J11	POSTGRADUATE DEGREE	5 and Above	Below 5	1 year and above

The results 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. 7 (63.6%) of the judges (J1, J2, J3, J4, J5, J7, and J10) had H. ND, B. Sc, B, Ed and other equivalent degrees as their highest educational qualification. 4(36.4%) judges on the other hand (J6, J8, J9, and J11) postgraduate degrees as their highest educational qualification. Table 2 also shows that 8 (72.2%) judges (J1, J2, J4, J5, J7, J9, J10, and J11) had five(5) and above years of Physics teaching experience, whereas 2 (18.2%) judges (J3, J6) had below five years of Physics teaching experience. 7 (63.6%) judges (J1, J2, J4, J7, J8 J9, and J10) had five (5) and above years of experience as Physics examiners at O'level, 4 (36.4%) judges (J3, J5, J6, and J11) had below five years of experience as Physics examiners. Finally, 10 (90.9%) judges (J1, J2, J3, J4, J5, J7, J9, J10, and J11) had been with their present students (the sampled students) for a period of 1 year and above, whereas only 1 (9.1%) judges (J6) had been with their present students for a period less than 1 year. These results demonstrated the adequacy of the sample of judges who participated in the standard setting.

Research Question 3: What are the outputs (cut scores) and consequences (percentage pass) of the three methods of standard setting for the PAT in the different schools and for the overall judges' panel?

Figure 1: Sub-panel and Overall Cut scores set for the PAT Using the Three Methods for the Different Schools



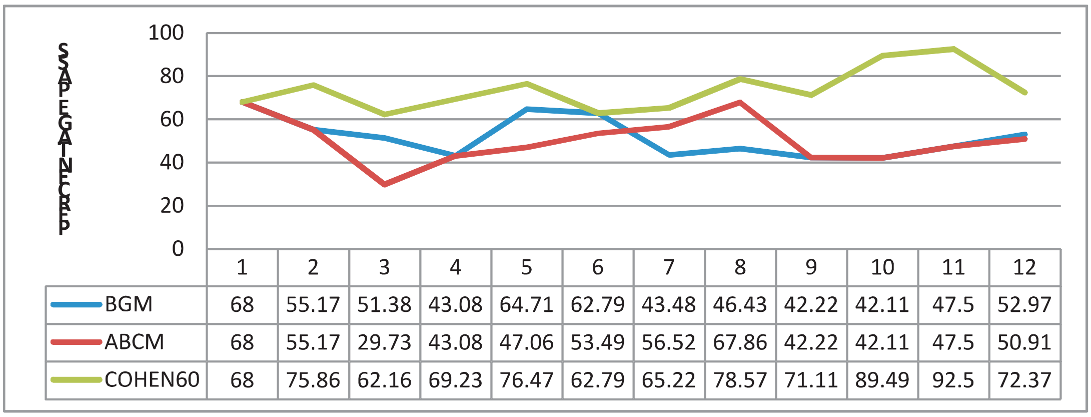
	1	2	3	4	5	6	7	8	9	10	11	12
BGM	19	23	16	19	16	17	17	37	20	22	33	19
ABCM	19	23	18	19	17	18	16	32	20	22	33	19
COHEN60	19	19	15	16	14	16	15	28	16	17	25	16

Key: BGM: Borderline Group Method; ABCM: Angoff/Borderline Compromise method; COHEN60: Cohen60 method.

Results presented in Figure 1 shows the individual cut scores set for the PAT using the three methods for the different groups of test takers. The Borderline Group Method (BGM), Angoff/Borderline Compromise method (ABCM), and Cohen60 method yielded cut scores which showed much similar fluctuation. However, the Cohen60 method can be seen to consistently give lower cut scores than the other two methods, except for school 1 where the three methods gave the same cut score of 19. The BGM and the ABCM apart from school 1 also produced the same cut for schools 2, 4, 9, 10 and 11. Other cut scores set by the two methods (BGM and ABCM) were quite comparable as seen from Figure 2 above.

Figure 1 also showed results of the overall cut scores set by the three methods using the data of the entire sample. The overall cut scores presented in column 12shows that the BGM and the ABCM produced the same cut score of 19, while the Cohen 60 method yielded a passing score of 16 three score scale points below the BGM and ABCM passing scores.

Figure 2: Percentage Pass Obtained in the Different Schools and Overall Sample Using Cut Scores from the Three Methods



	1	2	3	4	5	6	7	8	9	10	11	12
BGM	68	55.17	51.38	43.08	64.71	62.79	43.48	46.43	42.22	42.11	47.5	52.97
ABCM	68	55.17	29.73	43.08	47.06	53.49	56.52	67.86	42.22	42.11	47.5	50.91
COHEN60	68	75.86	62.16	69.23	76.47	62.79	65.22	78.57	71.11	89.49	92.5	72.37

Figure 2 shows the percentage pass obtained by using the cut scores set by the three methods for the 11 schools and for the overall sample. The ABCM cut scores produced the lowest percentage pass (29.73% for school 3) while the Cohen60 method gave the highest percentage pass (92.5% for school 11). Figure 2 also reveals that the percentage pass of the Cohen 60 method cut scores were consistently higher than those obtained from the cut score set by the BGM and ABCM except for school 1 where the three methods gave the same cut score, and consequently the same percentage pass.

Research Question 4a: Are there significant differences among the performance of the three methods in terms of Cut Scores?

Table 3: Difference in the cut scores set for the PAT by the three methods

ANOVA

DV: CUT SCORES

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	93.56	2	46.78	1.49	.241
Within Groups	1039.67	33	31.51		
Total	1133.22	35			

Table 3 reveals the difference in the cut scores set by the four methods for the PAT. From the table, the F-value, ($F_{(2, 33)} = 1.49$; $P > 0.05$) indicates that cut scores set for the PAT using the three methods differed significantly.

Research Question 4b: Are there significant differences among the performance of the three methods in terms of Percentage Pass?

Table 4 : Difference in Pass Rates Obtained as a Result of Cut Scores Set by the Three Methods for the PAT

ANOVA

DV: PERCENTAGE PASS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4121.90	2	2060.95	20.70	.000
Within Groups	3285.19	33	99.55		
Total	7407.09	35			

Result in Table 4 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, ($F_{(2, 33)} = 20.70; P < 0.05$) indicates that the percentage passes obtained from cut scores set for the PAT using the three methods did not differ significantly.

Table 5: Post-hoc test on the difference in pass rate obtained as a result of the cut scores set by the three methods for the PAT

Multiple Comparisons

Dependent Variable: PERCENT_PASS

Tukey HSD

		Mean			95% Confidence Interval	
		Difference	Std.		Lower	
(I) METHOD	(J) METHOD	(I-J)	Error	Sig.	Bound	Upper Bound
BGMCS	ABCMCS	1.35	4.07	.941	-8.65	11.34
	COHEN60CS	-21.99*	4.07	.000	-31.99	-12.00
ABCMCS	BGMCS	-1.35	4.07	.941	-11.34	8.65
	COHEN60CS	-23.34*	4.07	.000	-33.34	-13.35
COHEN60CS	BGMCS	21.99*	4.07	.000	12.00	31.99
	ABCMCS	23.34*	4.07	.000	13.35	33.34

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

The post-hoc results (Table 5) shows that the significant difference in Percentage Pass obtained among the three methods at $p < 0.05$ was as a result of the significantly high mean difference in Percentage Pass obtained between the Cohen60 Method and the other two methods (21.99 and 23.34 for BGM and ABCM respectively). As these were the only significant mean differences.

Research Question 5: To what extent are the cut-scores set by the three methods generalizable over the judges' panel used?

Table 6: Paired Sample t-test of Sub-group Cut Score Percentage Pass and Overall Cut Score Percentage Pass

		Paired Samples Test							
					95% Confidence Interval of the Difference				
		Mean	S.D	SEM	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	BGM - OVERALL_BGM	-6.13	30.35	9.15	-26.51	14.26	-0.66	10	.518
Pair 2	ABCM - OVERALL_ABCM	-7.41	24.62	7.42	-23.95	9.13	-1.00	10	.342
Pair 3	COHEN60 - OVERAL_COHEN60	5.79	30.54	9.21	-14.72	26.31	0.62	10	.543

Results presented in Table 6 shows that the percentage passes resulting from cut scores set by individual judges and the percentage pass resulting from cut scores set by using the overall cut score with the BGM across all the schools had a mean difference (X = -6.13; S.D = 30.35), which was found not to be significant with (t = -0.69; P > 0.05).With the ABCM the mean difference (X = -7.41; S.D = 24.62) was not significant too with (t = -1.00; P > 0.05), and with the Cohen60 the mean difference was (X = 5.79; S.D = 30.54) was also not significant with (t = 0.62; P > 0.05). The three methods therefore produced cut scores that are generalizable over the judges, panel used, since the consequences of the one-member sub-panels cut scores set independently are not significantly different from the consequences of using the overall cut scores set by all the judges, using the three different methods. The table also reveals that with (S.D = 24.62) being the least, the ABCM method used by the different judges produced cut scores that yielded percentage passes that were most consistent with the percentage passes from the overall cut scores for all the schools, hence the ABCM was the best performing method of standard setting in the study.

Discussion of Findings

The study compared the performance of three methods of standard setting for Physics Achievement Test (PAT) in secondary schools in Oyo state, in terms of the cut scores they produce, the consequences of the cut scores, and the generalizability of the cut scores over the judges overall panel used. Research question 1 aimed at providing statistical information on the performance of the students taking the test. It is seen from the variation of the mean, minimum and maximum scores that the performance of the students taking the PAT was widely varied. This wide variation according to Adewale and Antia, (2016) could be as a result of the fact that different teachers, with different characteristics taught the different student groups. Also the students may not be at the same level in terms of content coverage from where the PAT was constructed at the time of their exposure to the test. Consequently, varying cut scores should be expected with the use of norm-referenced methods of standard setting for each of the schools depending of the average performance of students within the school.

To demonstrate that the teachers that were involved in the process that will lead to the setting of such performance standards at the different schools (as one-member sub-panelists) are for the study, Research question 2 yielded results that indicated that all the 11 judges were adequately suitable to function as judges in the study. This is in line with Raymond & Reid, (2001) criteria for judge selection, which can be directly measured by the number of years of experience of the judges in; working with the examinee population, measurement and assessment, working with examination boards in a bit to ensure the validity of the process and whatever standards emerge from the process.

Research question 3 yielded results which indicated varying cut scores set by the Borderline Group Method, Angoff/Borderline Compromise method, and Cohen60 method for the different schools. And also varying consequences (percentage pass) resulting from the cut scores set. Comparing the variations in cut scores the results revealed that lower cut scores were set for schools for which the students had low performance in the PAT, and higher cut scores were set for schools for which the students had higher performance. This means that these methods are sensitive to group performance for a particular cohort (school). This agrees with the concerns and findings of Adewale and Antia, (2016), Bhandary (2011),Cohen-Schotanus and Van der Vleuten, (2010) and Searle (2000), Schoeman, (2011),that methods that are sensitive to such variations as test difficulty or group performance are better than the use of fixed passing scores.

Research question 4 allowed for results that revealed that the cut scores set by the three methods for the PAT did not differ significantly, but their consequences did. 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 generalizability of the performance standards is limited, and the validity of the performance standards is significantly reduced. The significant difference found among the percentage pass obtained when implementing the cut scores set using the three methods further confirms Kane, (1994) submission that, the impact or percentage pass, resulting from cut scores set are viable sources of validity information for the cut scores set using any method. The implication of this result is that though the cut scores set by different methods or different judges do not differ significantly, they could have very different consequences when being used to make decisions on students' performance in tests. A difference of 1 score point in test passing score could have grave impact on the adjudged performance of students in the test.

To further scrutinize the consequences of the cut scores set by the three methods in order to arrive at the most generalizable and defensible method for use with the PAT and similar testing situations, Research question 5 investigated the consistency of the consequences of the one-member sub-panels' cut scores with the overall panels' cut scores set using the three methods. The three methods were found to produce generalizable cut scores judged by the consistency of the consequences of the sub-panels' cut scores with the overall cut scores. However, the ABCM which tended to produce the highest mean sub-panel cut scores, produced cut scores that were also most consistent with the overall panel cut scores. The ABCM therefore stands out as the most generalizable and defensible standard setting method for the PAT with a passing score of 19/70 (27%)

Conclusion

From the results of the study, it can be concluded that the BGM, ABCM, and Cohen60 method produced varying sub-panel cut scores, but comparable overall cut scores. Scrutiny of the consequences of the sub-panel cut scores relative to the overall cut scores for the three methods identified the ABCM as the most generalizable and defensible method for use with the PAT.

Recommendations

- The ABCM should be adopted for use in the classification of test takers achievement in small scale testing programs such as school examinations,
- Public examination bodies and professional certifying or licensure examination in Nigeria should abandon the use of less objective and indefensible approaches to standard setting in favour of formal methods of standard setting such as the ABCM.
- More studies should be carried out on standard settings with attention focused on setting multiple cut scores with a variety of methods.

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. *Educational measurement*. Eds. R. L. Thorndike. 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.

Bhandary, S. (2011). Standard Setting in Health Professions Education. *Kathmandu Univ Med J (KUMJ),* 9, 3-4.

Bahry, L. M. Hambleton, R. K. Gotzmann, A. J. De Champlain, A. & Roy, M. (2012). National Assessment Collaboration Standard Setting Study Report. Prepared for the Medical Council of Canada while interning Summer 2012.

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.

Cohen-Schotanus, J. Van der Vleuten, C. P. M. & Bender, W. (1996). Een betere cesuur bij tentamens [A better standard setting method for written tests]. Gezond Onderwijs, 5(0), 83-88

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.

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

Hambleton, R. K. Jaeger, R. M., Plake, B. S. & Mills, C. N. (2000). *Handbook for setting standards on performance assessments.* Washington, DC.: Council of Chief State School Officers

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

Jaeger, R. M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice,* 10, 3-6.

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. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425 – 461.

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.

Norcini, J. J. Stillman, P. L. Sutnick, A. I. Regan, M. B. Haley, H. L. Williams, R. G. & Friedman, M. (1993). Scoring and standard setting with standardised patients. *Evaluation and the Health Professions,* 16, 322 - 332.

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.

Schoeman, S. (2011) Setting standards in health sciences education – a wake-up call. *African Journal of Health Professions Education* 3(1), 2 - 17

Searle, J. (2000). Defining competency–the role of standard setting. *Medical Education*, 34(5), 363-366.