Comparing Between Political and Standard-Based Cut-Scores in Muhammadiyah Vocational High Schools

Jahidatu Lis Silmi I'la Alhaq[1] and Haryanto[1]

[1]*Educational Research and Evaluation, Yogyakarta State University, Jl. Colombo No. 1, Yogyakarta, Indonesia*
alhaqsilmi@gmail.com, haryanto@uny.ac.id

Keywords: Achievements, Cut-score, Standard Setting, Vocational High School.

Abstract: Vocational high school in Indonesia need a factual standard setting method to ensure their students achievements. This survey intended to compare both standard setting based on scientific methods and political motives. Moreover, the researcher wanted to show the most usable standard setting to determine cut-score based on both item probability and students' abilities. The study targeted six Muhammadiyah vocationals high schools located in Sleman regency. All 10th graders were selected by using entire population sampling (239 students). To gather the data, the researcher collected all answer sheets from final exam of simulation and digital comunication subjects. Standard setting methods by Angoff and Ebel were compared to political cut-score used in Muhammadiyah vocational high schools. The findings that the cut-score currently used, 75, showed low level of Simulation and Digital Communication content mastery.

1 INTRODUCTION

Cut-score, sometimes called pass mark or standard score or pass point, is a special score that serves as boundary between those who perform well enough and those who do not. The latter is still a problem for many countries around the world. The academic meeting held at European Board of Ophthalmology, under the name CESMA meeting, Brussels, November 28th, 2009, had the same issue on its agenda. During the same meeting, the officials discussed different methods to be used in European countries for setting pass marks for examinations. The methods like Angoff, norm reference, contrast groups, Ebel, judgements, Nedelsky's, and among others, are the most used methods to determine passing score and they are absolute standards. However, some countries including Indonesia still use relative mark (norm-reference) for criterion-reference tests. Moreover, the passing mark in Indonesia is 75, and if a student does not answer 75 of all given questions, the teachers give unlimited chance until he/she reaches the minimal passing score. Using scientifically computed cut-score is a sustainable solution rather than determining it arbitrarily.

On a test, a passing score is a special score that is always used in order to determine the boundary for students who performed well and others who did not. With the broad context, this expression is considered as a professional value in testing the content and purpose, the test takers' abilities, and even the broad setting of education. The fact that standards are an expression of values means that a method for setting them is not a technique for divining a scientifically correct solution. Instead, it is a tactical mechanism to collect value judgements, achieving consensus, and explaining this consensus like a single point of a test. As long as the standards are judgement based, all methods of determining them will look similar in terms of their truth discerning ability. They will, however, vary in their credibility according to who sets the standards, the characteristics of the method they use, and the outcome (Norcini, 2003). To solve the problem of subjectivity, the cut-scores computed based on item probability are embroiled in this study and the results are compared to the never changed cut-score used by Indonesian vocational high schools.

It is very important to conduct this kind of study. Different institutions for licensure, education, and credential, are now in the sake of new methods for assessing professional competences. The recent commitment is centered on the will to find performance standards, which serve in separation of competent and non-competent examinees. The standard setting for assessing the performance is a new area for different but updated studies in education. As consequence, no one can recommend the best method for setting the passing score (Barman, 2008). The current problem is very important for Indonesian vocational high school's policy makers and other readers around the world. The big number of Indonesian vocational high schools do not even know how to determine cut-score for criterion reference test. By reading the current study, Indonesian academicians can have an idea on how to scientifically compute cut-score based on item probability, which is a result of examinees abilities.

In designing assessment tasks, test developers incorporate meaningful and essential performance criteria designed to provide evidence that candidates have successfully completed the task. Ideally, making the correct answer to the question in a test and achieving a maximum score are the dominant ways for candidates to demonstrate their mastery of the content taught. A requirement of mastery approach to performance (fully competent) for passing the test may appear unrealistic in most situations, owing to the complex nature of an educational task and measurement error (Ben-David, 2000). The researches like the current one are needed, especially in Indonesia, because some teachers only determine competent and less competent students based on the cut-point commanded by the government, 75 for Indonesian vocational high schools.

2 REVIEW OF LITERATURE

Standard setting is a critical part of educational, licensing, and certification testing. But outside of the cadre of practitioners, this aspect of test development is not well understood. Standard setting is a mechanism of defining the proficiency or achievement level and the cut-scores relating to those levels. A cutoff score is a point that helps to make classification of students with the objective of knowing the number of those in the group below the cut-score and that of those in the group above the cut-score, the class above the passing score is in the high level while the counterpart is in the low level. Frankly speaking, unless the passing points are not cautiously set, the outcome of the assessment become critical. This is why standard setting is a component of test development procedure which needs much attention (Bejar, 2008).

In a standard setting meeting using the analytical method (Hambleton, Jaeger, Plake, and Mills, 2000), panelists review a range of performance samples on all sections of an assessment individually. That is, if there is an essay and a multiple-choice portion of an assessment these portions would be reviewed and rated separately. Once the initial ratings of the first section are completed the panelists review ratings as a group and are given an opportunity to make changes to their ratings. Next, panelists rate the second portion of the assessment. Cut-scores for each section are summed to produce a cut-score for the overall assessment. While this method is straightforward, it does require many candidate performances samples, preferably near the cut-score to enhance reliability. As with many holistic methods, the preparation for the analytical method may be time consuming as well as the standard setting exercise. Therefore, it is advised to follow the steps and pay attention to the selection of panelists, but the problem may be how to know and what could be criteria to choose panelists.

A variety of standard-setting methods have been developed. Contrarily, the majority of methods work best depending on a specific item type, and thus it is highly recommended to match the test format with the suitable method that will be used or, at the very least, which methods will not be used. For instance, the modified method of Angoff (Angoff, 1971) has been used this long time for the determination of test cut-score, but it is suited to dichotomous items. As it is well described in (Hambleton, Ronald, and Plake, 1995), Hambleton and Plake tried to make some extensions to the modified Angoff method until the latter becomes a performance-based task within its application. The Body of Work method is a more recent method for setting cut-scores but is designed for assessments with more open-ended tasks and fewer dichotomous items (Morgan and Michaelides, 2005).

In many situations, cutoff scores are based solely on the judgments of subject matter experts. Several methods have been proposed over the years that involve expert judgments, and the Angoff method (Angoff, 1971) has generally been preferred and used most often (Hurtzand Auerbach, 2003). The method by Angoff deals with asking the judges to process with the probability estimation for last competent students to answer correctly to each item making up a test (Livingston and Michael, 1982). All the estimates by the judges are gathered in the form of proportions or percentages in order to be aggregated across the whole test and for every judge. The average across judges becomes the cutoff score for the exam (Hurtz and Hertz, 1999).

Assessment standards are relative/norm-referenced and absolute/criterion-referenced. Author called George tried to find a significant difference between assessing with norm and criterion referenced tests. Relative standard is used to identify a certain number of best examinees from the group. The first is simple in its administration but it is criticized by many to be not appropriate for authentic assessment. Norm-referenced standards do not indicate the competency relating to the job. However, another author called Hoover made a claim that this testing system (norm referenced test) is gives much information on the performance of examinees than criterion referenced test can do. Most examining bodies use norm-referenced system in reporting students score. When the primary objective of medical education is to create a society with skilled physicians, the assessment is expected to focus on the difference between competent and non-competent and not rank order of individual prospective physicians in the group. In educational systems around the globe, the criterion referenced tests are used to assess students against the fixed objectives or indicators of achievement that are in the syllabus and distinguish students with high level of competence from those with low level of competence. The criterion standard gives the information about every examinee's degree of competence without depending on the performance of his/her fellows in the same group. The test should be in the same line with the goals of the course supposed to be delivered to students and against all standards and criteria that are already pre-specified. The majority of assessments that involve high-stake competencies always require absolute standards for passing or failing decision. The criterion reference tests are developed in such way that provides with a direct meaning to the measurement in terms of a chosen standard of performance. The same type of testing system has the standards of performance within itself. The term criterion has the same meaning as standard or cut-off (Barman, 2008).

Some researchers and educational departments empirically studied concerning standard setting. American Educational Research Association, American Psychological Association, & American Council on Measurement in Education (1999) suggest several soundness criteria, such as when proposed score interpretations involve one or more cut-scores, the rationale and procedures used for establishing cut-scores should be clearly documented. The comments that came after emphasize that the appropriate decision in every score scale region at which the passing points are supposed to be strictly requires reliable techniques of classifying the examinees into categories. An advanced standard in determining the significance of the various classifications, particularly the proficiency designation, needs an audit or empirical comparing task with an external test (Hamilton,McCaffrey, and Koretz, 2006). A couple of previous reports used that approach in order to examine the degrees of proficiency across states against national benchmark. The findings out of the both studies come up with the proof of a remarkable difference among all states in terms of the student's proficiency designation proportionality (Bejar, 2008).

The study by Braun and Jiahe(2007) used the National Assessment of Educational Progress (NAEP2) as the common yardstick for comparing states' proportions of students classified into the groups of mathematics and reading competency against the NAEP findings about each state. NAEP only worked on reading and mathematics as long as all states do so during NCLB test, but NAEP always pushes out its own definition of degree of proficiency and its unique approach for both mathematics and reading assessment, which is different from the approach adapted by each state. For example, NAEP puts interest in a significant partition of items with response construction as a primary requirement, that is, the questions in a test oblige the examinees to provide the answers by their own, such as essays or blank filling questions, instead of selecting the correct alternative from a list of options. The conclusion in the reports stated that the degrees of student's achievements differ depending on how each state defines the word proficiency, hence, that is their specific passing points to determine the achievement level of their students (Braun and Jiahe, 2007). To find the dissimilarity across the achievement levels is dependent is not necessarily dependent to the education quality change particularly in its system from state to another.

There are questions to find answers to in this study. With the results out of this work, the following questions should be answered:

a. How accurately do cut-scores reflect to student's achievement?
b. How does passing mark 10th graders from six Muhammadiyah vocational high schools located in Sleman regency?
c. What are the standard setting methods match the test formats used in Muhammadiyah vocational high schools across Sleman regency?

3 RESEARCH OF METHOD

This study involved quantitative approach and it specifically stressed on cross-sectional survey. The study was called so because the data were collected to make inferences about population of interest at one point in time. The variables in the study are achievement level of students in Simulation and Digital Communication subject.

Concerning the description of the participants, the study targeted six Muhammadiyah vocational high schools located in Sleman regency. The regency of Sleman is located in Special Region of Yogyakarta (DIY). The participants, 10th graders, were selected purposely. Total participants are 239 students from SMK Muhammadiyah Cangkringan, SMK Muhammadiyah Berbah, SMK Muhammadiyah Minggir, SMK Muhammadiyah Mlati, SMK Muhammadiyah Gamping, and SMK Muhammadiyah Seyegan. Item response theory was used to calibrate students answers for the first semester Simulation and Digital Communication test 2018/2019.

4 FINDINGS

For the findings about students' scores on Simulation and Digital Communication test for first semester 2018/2019, the researcher made central focus on central tendencies descriptors. The consideration involved the mean to show the central score, median to show the mid-score if all examinees scores are ordered, mode to see the score that many examinees have, and standard deviation to see how all examinees scores disperse from the mean score. Therefore, the descriptive statistics indicated that M = 52.72, Median = 48.98, Mode = 46.94, and SD =16.72. The descriptive statistics demonstrated that 10th graders' performance in Simulation and Digital Communication subject is still low for six Muhammadiyah vocational high schools located in Sleman regency.

Concerning pass mark, the researcher based on two mostly used standard setting methods. As discussed before, all methods are almost similar; the study involved the widely used methods. A big number of researchers frequently use the methods for standard setting founded by Angoff and Ebelto determine the pass mark. For each method output, the researcher tried to show number of students who reach the pass mark and those who did not.

Table1: The Angoff's Standard Setting Method Results

School	Cut score	N-Pass	%	N-Fail	%	Tot
Cangkringan	61.64	56	51.38	53	48.62	109
Berbah	40.29	28	59.57	21	44.68	47
Minggir	37.73	9	45.00	11	55.00	20
Mlati	44.92	14	58.33	10	41.67	24
Gamping	47.16	17	62.96	10	37.04	27
Seyegan	58.07	7	50.00	7	50.00	14
Overall	48.24	131	54.81	108	45.19	239

Table 1 contains information about students' achievement in Simulation and Digital Communication subject. Based on Angoff's standard setting method, each school out of six schools has its own cut-score. For SMK Muhammadiyah Cangkringan, the cut-score is 61.64% and only 56, (51.38%) students out of 109 passed the mark while 53 students, (48.62%) did not. For SMK Muhammadiyah Berbah, the cut-score is 40.294% and only 28 students (59.57%) passed while 21 students, (44.68%) failed. For SMK Muhammadiyah Minggir, the cut-score is 37.732% and just 9 students (45%) passed whereas 11 students (55%) failed. For SMK Muhammadiyah Gamping, the cut-score is 44.92% and 14 students (58.33) while 10 students (41.67%) failed. The cut-score for SMK Muhammadiyah Gamping is 47.164% and only 17 students (62.96%) passed whereas 10 students (37.04%) failed. The cut-score for SMK Muhammadiyah Seyegan is 58.072% and just a half of students (7 = 50%) passed. The pass mark for all schools is 48.241% and 131 students (54.81%) passed against 108 students (45.19%) who failed. To make it clearer, the figure 1 illustrates the overview of Table 1.

Table 2 gives the information about the student's achievement based on Ebel's standard setting method. The method always relies on the item difficulty, easy, medium, and hard. As Mardapi (2012) described, an item is classified in a medium category if its difficulty indices vary from -2 to 2, easy if difficulty indices are less than -2, and hard if difficulty indices are bigger than 2. As it is illustrated in the table, most items are classified in medium category, 43 items (86%). After computing the expected score for each category, the pass mark is 52.44% for all schools located in Sleman regency. The number of students who passed is 107 out of 239 students (44.77%) while a big number of 132 students (55.23%) failed. Therefore, the rate of failure outweighs that of success among all 10th graders from six Muhammadiyah vocational high schools in Sleman regency. Table 2 centers on the information on figure 2.

Table 2: The Ebel's Standard Setting Method Results

Item Difficulty Category	Number	Probability correct	Expected Score
Easy	3	1	1*3=3
Medium	43	.535	43*.535=23.22
Hard	4	0	4*=0
Sum			26.22
Cut-score (26.22*100)/50			52.44
	Decision		
N-pass	%	N-Fail	%
107	44.77	132	55.23

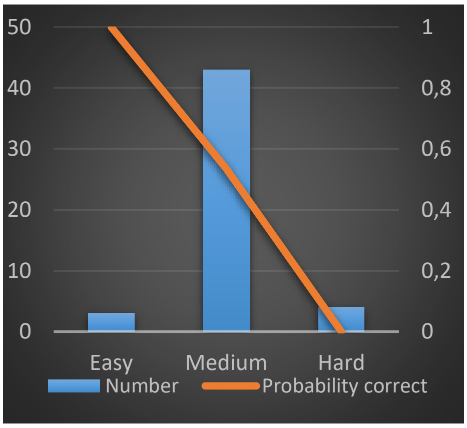

Figure 2: Item Difficulty and their Corresponding Probability

Table 3: Students Achievement Based on Indonesian Cut-score (75)

School	N-Pass	%	N-Fail	%	Tot
Cangkringan	32	29.36	77	70.64	109
Berbah	0	0.00	47	100.00	47
Minggir	3	15.00	17	85.00	20
Mlati	1	4.17	23	95.83	24
Gamping	2	7.41	25	92.59	27
Seyegan	3	21.43	11	78.57	14
Overall	35	14.64	204	85.36	239

Table 3 englobes the information about the effects of unrealistic or political pass mark. This 75 used in Muhammadiyah vocational high schools does not base on students' abilities, item characteristics, or item probability. It is clear that for all schools the number of failures (<75) outweighs that of success (>75). There is one school, SMK Muhammadiyah Berbah, where there is no success, (0%). For all schools together, only 35 students (14.64%) passed while 204 students fail, (85.36%). Hence, the level of achievement in Simulation and Digital Communication subjects is completely low comparing to how it would be if Indonesian policy makers reflect on standard setting methods. To completely display situation in table 3, figure 3 explains more.

5 CONCLUSION

This study intended to compare the level of student's achievement, on Simulation and Digital Communication test, by relying on both scientific and unrealistic or political cut-scores. The researcher also wanted to see whether there would be impact of having a fixed cut-score for different levels of student's abilities. Item response theory is one of the best ways to estimate students' abilities in order to avoid victimizing examinees when wanting to determine the number for promotion or repeat. After getting the finding, it was proved that six Muhammadiyah vocational high schools located in Sleman regency, there is mismatch between students' scores on Simulation and Digital Communication test and the cut-score always relied on for determining their levels of achievement. Therefore, students are victims of the fixed cut-score (pass mark 75) used in Muhammadiyah vocational high schools.

6 SUGGESTIONS AND RECOMMENDATIONS

According to the findings out of this study, the researcher is suggesting that:

a. The government should set the pass score after equating students' scores per localities;

b. Academics should determine the pass mark based on students' abilities, not only for Simulation and Digital Communication subject but also other subjects.

c. Teachers should avoid forcing until a lowly able student reaches 75 to pass because such student may fail in the final examination.

REFERENCES

Angoff, W.H. (1971). *Educational Measurement*. Washington DC: American Council on Education.

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

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

Ben-David, M.F. (2000). AMEE guide no. 18: Standard setting in studentassesment, *Medical Teacher, 22* (2), 120-130.

Braun, H.I. and Qian, J. (2007). *An enhanced method for mapping state standards onto the NAEP scale. In Linking and aligning scores and scales*. New York: Springer.

Cunningham, J. (2012). *Student achievement. In Washington DC*. US: National Conference of State Legislatures.

Hambleton, R.K. and Plake, B.S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments, *Applied Measurement in Education, 8* (1), 41-55.

Hambleton, R.K., Jaeger, R.M., Plake, B.S. and Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24* (4), 355-366.

Hamilton, L.S., McCaffrey, D. F., and Koretz, D. (2006). *Longitudinal and value-added modelling of student performance. In validating achievement gains in cohort-to-cohort and individual growth-based modelling contexts*. MN: JAM Press.

Hurtz, G.M. and Hertz, N.R. (1999). How many raters should be used for establishing cut-off scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement, 59* (6), 885-897.

Hurtz, G.M. and Auerbach, M.A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cut-off scores and judgment consensus. *Educational and Psychological Measurement*, *63* (4), 584-601.

Mardapi, D. (2012). *Pengukuran, Penilaian, dan Evaluasi Pendidikan*. Yogyakarta: Nuha Litera.

Livingston, S.A. and Zieky, M.J. (1982). *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton: Educational Testing Service.

Morgan, D.L. and Michaelides, M.P. (2005). Setting Cut-scores for College Placement. Research Report No. 2005-9. *College Board*.

Norcini, J.J. (2003). Setting standards on educational tests. *Medical education*, *37* (5), 464-469.