Journal of Research and Reflections in Education
June 2007, Vol. 1, No. 1, pp 66 – 87
http://www.ue.edu.pk/jrre

Criterion Referenced Setting Performance Standards with an Emphasis on Angoff Method

Muhammad Naveed Khalid and Muhammad Saeed

Abstract: This paper explains the concept of standard setting and the different methods used in determining the cut-off points of several examinations for certifying or licensing a candidate with an emphasis on Angoff Method. The standard setting methods are useful tools for making informed judgments. No single method for standard setting is suitable for standard setting in different situations; each method works best with a particular item type, and thus matching the test format to an appropriate method helps to determine which standard setting method will be used? The most popular method is the Modified Angoff method, which is typically used to set standards for tests with primarily multiple-choice or dichotomous items, does not truly include open ended questions. The Body of Work, on the other hand, has an edge of being more effectively used for open ended task, but it deals with a fewer dichotomous items. There is a need for further research in the consistency and replication of results by using different subjects, judges and situations. The findings may be beneficial for the Higher Education Commission Pakistan to introduce criterion-referenced standard setting in the higher education institutions in its various programmes.

Keywords: standard setting, criterion-referenced, norm-referenced, cut-off scores

Introduction

One of the key components of the educational process is measurement and evaluation of the learning outcomes. These help the educational planners, managers and teachers to make key decisions about learning and teaching. One of these areas is setting performance standards in order to determine the current status or academic performance of students. Many scholars define standard setting in different ways but all focus on one idea, that it involves judgment on the student's performance. As Linn (1979) stated the analogy between standard setting and legal practice for it requires many judgments as judges in the court room. His words as put by Cizek (2001, p.6), "the cut-off score

problem is very similar to one judges and lawyers deal with all the time; the question of where and how to draw the line." That is every one involved in the judgment process should have a reasonable justification for defining the cut-off score.

Hambleton and Plake (1998) view standard setting as "the Achilles heel" of educational testing, because, as Kane (1994) perceives that "there is no clear consensus on the best choice among the numerous methods and the results of applying any method cannot easily be validated". Cizek (2001, p. 5) states that 'standard setting is the branch of psychometrics that blends more artistic, potential, cultural ingredients into mix of its products than any other'. Stephenson et al. (2000) describe standard setting in the context of pass/fail of students, as they assert 'standard setting is a process by which test scores can be classified as pass/fail, master/non-master, certified/not-certified and licensed/not-licensed.

Standard setting is usually a technical procedure or at least involves technically trained specialists, due to the requirement of establishing a cut-score on the test scale. It is an organized system for collecting the judgments of qualified individuals about the level of knowledge and skills needed for someone to be classified as above a standard. Standard and standard setting are not the same terms; sometimes people are confused with. Clarifying distinction between both of these terms, as Cizek (2001) states 'standard is the result and the standard setting is the method or means for achieving the result'.

Stephenson et al. (2000) incorporate the views of Zeiky (1994) who divided the rise of standard setting into four periods: age of innocence, awakening, disillusionment and realistic acceptance. The period of innocence captures the years before 1950. During this period a little or no attention was paid to how standards were set? When ability tests began to be used to classify people during World War I, there was a need to standardize tests for a large population. The growth of criterion referenced testing had a tremendous impact on the history of standard setting. The next period, the age of awakening, occurred with the rapid growth of criterion referenced testing and the minimum competency/basic skills testing in the early 1970s. The age of awakening, with its development of systematic methods of standard setting, was quickly overlapped by the age of disillusionment in the late 1970s. The period of disillusionment was the period during which researchers compared the methods and realized that the various methods

produced different cut-scores. The period of disillusion slowly dissipated into the current period, "the age of realistic acceptance".

The distinction between performance standard setting and cut-score setting is often confusing. Kane (1994, pp. 425-461) distinguishes performance standard and a cut-score as performance standard is the minimally adequate level of performance for some purpose, and a cut-score as a point on a score scale. Ricker (2006) states "standard setting is a philosophical and policy making activity, while setting a cut-score is the operationalization of that policy". In any way a cut-score is the score on the test or assessment chosen to select or classify examinees with respect to the performance standard. It is the score that is claimed to distinguish between those who have satisfied the performance standard and those who have not. Standard setting should fairly and accurately differentiate between different levels of student performance and must be understood by all stakeholders. For example, if a cut-off point of a course is certification of competence; then, it is important that each cut-off point accurately reflects a student's mastery of course materials, which depends upon the strategies or the categories (student-centred, examinee centred etc.)

It is also necessary to mention here that decisions relating to setting standard and cut-score are not free of political dimension. In other words, these may involve political decisions, as one purpose of standardizing test is to increase "accountability". There is an increased pressure on the people who set out cut-scores to reduce legal vulnerability and to maintain fairness. In addition, to set a standard is to develop a policy and policy decisions are not right or wrong. They can be wise or unwise, effective or ineffective, but they can not be validated by comparing them to some external criteria (Zeiky, 2001). There are no best ways to set performance standards, although everyone agrees that standards provide information on how well students are learning? They are intended to convey the level of achievement of each student for the courses attended and later to be an overall accomplishment for the program he or she has studied.

Norm Referenced and Criterion Referenced Standard Setting

Performance can be defined either in relative or absolute terms by comparing students with each other or measuring their achievement against a pre-determined scale.

Some relative standard setting schemes make it impossible for the students to estimate their final grades because the cut-off points in the final distribution are not determined until the end of the course. Some kinds of comparison are made when performance standards are assigned. For example, a teacher may compare a student's performance to that of his or her classmates, to standards of excellence, i.e. pre-determined objectives, performance indicators or to combinations of each.

Norm Referenced Setting Standards (NRSS) is based on comparisons of students' achievement with other students. By comparing a student's overall course performance with that of some relevant group of students, the evaluator sets the cut-off point to show the student's level of achievement or standing within that group. In this form of setting standard an "A" (in A, B, C, D, and F scaling where A is the highest) might not represent *excellence in attainment of knowledge and skills* if the reference group as a whole is somewhat incompetent. The nature of the reference group used is the key to interpret the ranking based on comparisons with other students. Comparing students' performance with others, also called norm-referenced scaling system, and this is based on a pre-set distribution of scores. The benchmark for each grade level varies across borders or even in different institutions in the same country. Generally, this standard setting is taken around: 'A' top 10%; 'B' next 20%; 'C' next 20%; 'D' next 20%; 'E' next 15%; and 'F' bottom 15%. Using such group comparisons for defining the levels are appropriate when the class size is sufficiently large to provide a reference group representative of students enrolled in the course (Garcia-Quintana and Mappus, 1980).

Criterion Referenced Standard Setting (CRSS) is based on minimum set criterion or established standards. Here the cut-off point is obtained by comparing a student's performance with specified absolute standards rather than with such relative standards as the work of other students. In this method, the teacher is interested in indicating how much of a set of tasks or ideas a student knows, rather than how many other students have mastered more or less of that domain? A "B grade" in a course might indicate that the student has 'average or good or just satisfactory level of competency'. Here the teachers set the minimum criterion in regard to students' learning achievement and assess them on that criterion. There is a possibility that a large number of students

would achieve that criterion or just a few would achieve it. Comparisons are not made with their classmates or students of the same grade levels of other sections or schools. A group of students achieving ‘grade A’ might not able to achieve ‘grade B’ on tests designed by other teachers in the same school or other schools for the same class (Berk, 1986).

The key concern to developing CRSS is to clearly define the course goals and standards before their assessment and evaluation. For the standard-based grading to be effective, course goals and standards must necessarily be defined clearly and communicated to the students. Most students, if they work hard enough and receive adequate instruction, can obtain high scores. The focus is on achieving course goals, not on competing for a grade. The performance level of a student indicates "what" a student knows rather than how well he or she has performed relative to the reference group? Table 1 displays the assumed scores range achieved by students (first column), percentage of students met this criterion (second column), and performance level (last column) indicating achievement of students. There are two main drawbacks of CRSS: First it is time consuming, and second it is not easy for each teacher to truly determine what course standards should be for each possible course grade issued? (Saeed, 2002).

Grade compares performance either to a relative standard (norm-referenced) or to an absolute standard (criterion-referenced). Norm referenced tests are designed principally to facilitate the use of scores derived from the tests to make comparative statements about individuals. This is not the primary type of information required by teachers who implement objectives based on programs. They require information about the level of individual performance relative to well-defined content domains (Hambleton, 1978). The key purpose of NRT is ‘to rank students performance with their classmates or other students of same grade level’ (Saeed, 2002). For example, a relative comparison is being made if, let’s say, a “C” grade means "average performance compared to others in the class," but an absolute comparison is being made if it means" demonstrated attainment of the most important objectives". In other words, if a student or group of students achieve four out of five ‘mastery level’ in a certain skill, we would interpret the results that the individual student or the group has attained the desired mastery level and all students attaining the four of the five ‘master level’ will be placed in one group. It is

essential for an institution or department to adopt a standard based or criterion-referenced meaning to develop a description of the learning outcomes that define each grade symbol. Table 2 illustrates the types of expressions that can be used to differentiate levels of performance on the absolute and relative grading scales.

Charry and Ellis (2005) in regard to the comparison between the two approaches (NRT and CRT) based on their experimental design concluded that norm referenced (rank-order) grading can generate improved student performance relative to a criterion-reference grading system. While such norm-reference systems may not be appropriate in all settings, our results suggest the method works well in university, principal-level courses. Indeed, the attributes of principal-level courses create an attractive setting for rank-order grading: (1) relatively large enrolments and (2) no explicit student co-operation. Criticism on norm-referenced methods cites the potential for competition to harm the education process, but competition may be its best virtue. Students respond to incentives and the stronger incentives arising from competition can motivate improved student performance, especially among high performing students. But in the wrong setting, competition may inject negative aspects to the learning process. The decision to use rank-order grading should consider the positive and negative impacts and the decision will differ across different educational settings.

Standard Setting Methods

At present, there are many standard setting methods developed by different scholars for a number of situations that do require decisions. These methods have their own strategies or some what unique requirements to fulfill the type of administrations test based or group dependent, and the different requirements for selecting and training participants. Literature cites different standard setting methods (Measurement Research Associates, 2007; Vos, 2004; Morgan and Perie, 2004; Stephenson et al., 2000) based on their judgment on the contents or the results obtained from the test.

Method Based on Judgment about Individual Test Takers

The methods based on judgments about test takers require two types of information about each test taker: The person's test score and a judgment of the adequacy of the test taker's knowledge and skills. The judgments used in these methods

should meet the four requirements. First, judgments must be made by qualified persons. The judges must be able to determine each test taker's knowledge and skills and must know what level of knowledge and skills a person passing the test should have? Second, judgments must be based on the skills and knowledge that the test is intended to measure. The judges must understand which characteristics of the test takers they should judge and which they should disregard. Third, judgments must reflect the test takers skills at the time of testing. Fourth, judgments must reflect the judges' true opinions. It is important that the judges should be objective in their judgment (Stephenson et al., 2000).

Contrasting Groups' Method

The principal focus of judgment in the contrasting groups' method is on the competence of examinees rather than on the difficulty of a test or its items. The contrasting-groups method is based on the idea that the test takers can be divided into two contrasting groups, a qualified group and unqualified group, on the basis of the judgments of their knowledge and skills. As obvious choice for a passing score in this case would be the score at which there is just as many qualified test takers as unqualified test takers, (Livingston and Zeiky, 1982). A variation of this method involves asking judges who have knowledge of both the examinee and the required knowledge or skill level to classify examinees into one of three categories: competent, borderline, or incompetent, and the standard is based on analysis of the test score distributions of examinees who are judged to be competent or incompetent. To come up with a passing score, Stephenson et al. (2000) recommend that the two score distributions be plotted and the point of intersection of the distributions be chosen as standard. This method has two main advantages: First is the ability to accommodate both dichotomously scored and polytomously scored items, and the second is the ability to collect the data prior to the administration of the examination (Skakun and Kling, 1980). Contrasting groups' method is considered a good method to use when revisiting cut-score decisions to provide confirmatory evidence that the decisions are still valid (or evidence of the need to run a new standard setting workshop). One disadvantages of this method is that it can be subject to how well panelists know students being classified and any personal feelings they have towards those students? (Morgan and Perie, 2004).

The Borderline Method

The borderline group method focuses on the qualifications of examinees, rather than on test items. Judges must possess to designate him/or her as competent. After a sample of judges has been selected, they are asked to categorize examinees into three categories: competent, borderline and incompetent. That result from the borderline group procedures, as Stephenson et al. (2000) state is that, 'the median of the distribution of test scores earned by examinees that are classified as borderline'.

Methods Based on Judgments about Test Questions

All standard setting methods are based on the idea that test takers who belong to the upper group tend to earn higher scores than those who belong to the lower group, the passing score should be the score that would be expected from a person whose skills are on the borderline. The judgments can be applied, either before or after the test is administered and requires judgment.

The Nedelsky Method

The Nedelsky (1954) method of standard setting is developed for multiple choice items. This method involves assigning values to test items based on the likelihood that examinees rule out incorrect options and then choose from the remaining options *(*Stephenson et al., 2000). Each judge should specify a response option of a student who is on the border of the expected qualification "sufficient/insufficient" must be able as being wrong. If such a borderline chooses at random between the remaining alternatives, the probability to answer the item correctly is equal to the reciprocal of the number of alternatives (Vos, 2004*).* Accordingly, the cut-off point for each judge is now established as the sum of correct probabilities across all items in the test and the mean or median of all judges' cut-off points can be used to determine the cut-off point for the subject matter.

The limitation of this method is that it only focuses multiple choice questions (MCQs), while usually the actual tests are not just based on MCQs. Moreover, the borderline students may have partial knowledge on the alternatives which makes the assumption unrealistic that the students know all wrong alternatives and choose randomly (Vos, 2004*).*

The Bookmark Method

In the Bookmark method test items are ordered from easiest to most difficult based on Item-Response Theory (IRT) *b*-values, difficulty parameters. Panelists are asked to consider items in the order of difficulty and identify the place in the ordered item booklet where the borderline students at each performance category would have a specific probability, traditionally two-third 'response probability' (2/3rd RP means for a given cut-score, a student with a cut-score at that point will have a .67 probability to answering an item correctly) of getting the item correct. Panelists are instructed to place a bookmark into the ordered item booklet at the identified spot to mark their recommended placement for the cut-score. After three rounds of bookmark placement with discussion between each round, final round panelists bookmark placements are compiled and the median selected for the cut-score recommendation. This cut-score recommendation is then located on the IRT ability metric to find the place where students have a two-third (or other probability being used) chance of answering the identified item correctly and this becomes the final cut-score recommendation. Thus the RP adjustment is used both in the instructions given to panelists and in scaling the items. An advantage of the Bookmark method is the ability to set multiple cut-scores simultaneously. The method is also very efficient in terms of time needed and seems to be easily understood by panelists. This method works well with both dichotomously and polytomously scored items. However, a criticism is in the use of the RP67 value which can be confusing to panelists and authoritative bodies who think the panelists' bookmark placement (i.e., number of items preceding the bookmark) is directly translated as the recommended cut-score (Mitzel et al., 2001; Lewis et al., 1996) as quoted by Morgan and Perie (2004).

Angoff Method

The Angoff (1971) method is the most basic form of the criterion based setting standards perhaps due to relatively simple process of determining the cut-off points. Like in the Nedelsky method, judges in this method are expected to review each test item and passing score is computed from an estimate of the probability of a borderline candidate answering each item correctly. After discussion and consensus of the characteristics of a borderline candidate, each judge makes an independent assessment of the probability that a borderline candidate will answer the item correctly for each item. The judges'

assessments of an item are averaged to determine the probability of a correct response for that item. Then, each probability assigned to an item on the examination form is averaged to obtain the pass point (Measurement Research Associates, 2007).

Berk (1986) quotes following nine advantages of Angoff method, that's why it is more widely used in the world;

1. It yields appropriate classification information.
2. It is sensitive to student performance.
3. It is sensitive to instruction and training.
4. It is judged in the measurement literature to be statistically sound.
5. It takes measurement error into account.
6. It is easy to compute.
7. It is generally easy to explain to laypeople.
8. It is generally credible to laypeople.
9. It can be applied to many item formats.

Angoff method has two disadvantages: First it assumes judges to have a good understanding of the statistical concepts, and second is that panelists may lose sight of the students' overall performance on the assessment due to the focus on individual items, as this method carries item-based procedure.

Cizek (2001) quotes Raymond and Reid in regard to selecting and training participants for standard setting. As can be seen in Table 3, the relationship between selection and training as they apply to standard setting based on the idea of Angoff method. The first column identifies the major tasks required of participants during the standard setting study whereas the second column identifies some of the 'knowledge, skills and abilities' (KSAs) required to completing these activities. In other words, the first column specifies what standard setting participants need to do? The second column shows KSAs; the third and fourth columns list the selection and training activities that should be considered to assure that participants possess the requisite KSAs prior to providing standard setting judgments.

Plake et al. (2000) investigated the technical quality of results from Angoff's Standard Setting Method. In the context of role of reliability and validity in standard

setting, as they believe that for granting a license in their profession or certifying students from their school, minimum passing scores (MPS) are frequently used to make critical decisions about individuals based on their performance on the test. Therefore, the procedures set in defining the cut-off score need to be scrutinized or investigated for their validity and reliability. The Angoff method relies on experts/panelists making item performance estimates for minimally competent candidates (MCCs). The item performance estimates are aggregated across items and averaged across panelists to yield the recommended cut-score. Therefore, the accuracy and consistency of these items performance estimates is central to the validity of the resultant cut-off score.

Comparison of Angoff Method with other Methods of Establishing Cut-off Scores on CRT

Criterion-referenced tests are widely used to monitor progress through objectives-based instructional programs, diagnose student weaknesses, evaluate programs, and assess competencies on certification and licensing examinations in Hambleton and Egnor (1979) as put by Mills (1983). Most often, these tests are used to sort examinees into categories or states based on their performance on the tests. Any time examinees are to be classified into mastery groups based on test performance, a performance standard or cut-off score must be established. The main reasons for differences across these methods are in regard to procedures used in the data analysis regardless of the type of methods employed - judgmental or empirical.

Hambleton and Egnor (1979) found that the standard setting methods used for the comparison are the Angoff method, the contrasting groups' method and borderline method. Twelve test forms were piloted: Form A–F contained language arts items (a total of 162 items) and Forms G-L contained mathematics items (a total of 260 items). The first 10 items of all forms within a subject area were identical. Thus, Forms A-F all contained the same 10 common items. A 20% sample of second grade students was selected to participate in the pilot. The sampling design called for responses of approximately 1000 examinees to each test form. It was found that the Angoff and the contrasting groups' methods were almost similar. The contrasting groups quadratic discriminate (QDF) method sometimes produced results that were similar to those produced by the Angoff and contrasting groups (graph) technique in language arts. In

mathematics, however, the method produces a cut-off score of zero for five of the six tests. The results from the borderline group method differed Angoff and contrasting groups' method. The study resulted in congruent results from different standard setting methods for several test forms. It was found that the congruence was high likely because they used the same judges of different methods.

The Angoff method has recently come under criticism for being potentially invalid due to concerns about the ability of panelists to make accurate item performance estimates, especially for difficult or easy items. But this method has main advantage of that it is relatively straightforward, and does not require examination data. Here the experts derive judgment on their experience rather than rely on probability model like IRT (Dichotomous Model). For example, the Nedelsky method (1954) requires panelists to make judgments about the answer choices in multiple-choice test items. Because most examinations are not limited to employ solely the multiple-choice format, the methodology chosen needed to accommodate open-ended items as well. The Angoff method (1971) accommodates various item formats, and has been used in standard-setting initiatives with both multiple-choice and essay type questions (Stephenson et al., 2000; Vos, 2004*)*. Vos (2004) added that unlike the Nedelsky's method, here judges are required to explicitly specify the probability of the correct alternative for a borderline candidate rather than deriving this probability implicitly from behaviour on the distracters under the model of random guessing.

Hambleton and Plake (1979) explained Angoff's method as a direct extension of dichotomously scored multiple-choice tests, and was also used for setting performance standards. The procedure is based on the following three steps:

1. The judges are asked to provide passing standards on the dimensions used in scoring each exercise for a just barely certifiable (JBC) candidate on a 4-point scale ranging from substantially deficient to outstanding.

2. The judges are asked to assign weights to each dimension: within-exercise, reflecting their views of the relative importance of the dimensions. Doing so, the judges are told that these relative weightings will be used for computing the exercise standards and summing.

3. In addition, the judges are asked to assign relative weights to the exercises in order to make certification decisions. These assessment package standards were finally obtained by attaching the specified relative weights to the exercise standards and summing (Vos, 2004).

Conclusions and Recommendations

In view of the preceding discussion, it can be concluded that no single method is appropriate under all situations. The concept of standard setting and the different methods or approaches are used in determining the cut-off points of several examinations for certifying or licensing a candidate. This decision making process requires reliable data that helps judges to come out an informed decision. The standard setting methods, therefore, are the means, not the ends, so as to judge wisely. These help the judges to know more about what they are doing and how they can do it.

The scholars have also classified them into two as person or population based sometimes empirical based methods and test or item based standard setting methods. This way of classification came as a result of the new movement of the 1970s of criterion referenced setting standards (Cizek, 2001). At present there are promising standard setting methods by using the concept of item response theory (IRT) by maximizing the information at the cut-off score. Of all the most predominantly used standard setting method is the Angoff method. This method besides its critics for "difficulty and confusing" features, it is the most popular and widely used way of setting performance standard in licensing companies and schools. It is not only used in large or national examinations or it is not only applied so as to determine the minimally competent candidates, but could also be used to create a cut score for any grouping within the population too (Ricker, 2006). For example, we can use it to set cut-off score for a standard of excellence on a test. In this case judges would be required to conceptualize a student based on the scaling measures such as conceptualizing an A, B, C, D and F students' performance in going through each item on the test.

Which method should be used for assessing students' ability at higher level is an important question? Perhaps CRT has edge over NRT, as it can truly assesses the individual students' achievement and also looks into how the students meet the

minimum criterion over a period of time in attaining or improving certain body of knowledge or skills? The GRE type test in the context of Higher Education Commission (HEC) Pakistan is focused on norm referenced standard setting. The universities and colleges are basically offering programs and courses that are more or less similar. But different teachers are teaching these courses at different areas and time. When it comes to the cut-off score to these courses the teachers use their own way of grading them and the cut-off point is different for different teachers teaching the same course. These differences bother the newly authorized body for quality of education and accreditation. Different teachers use their own discretion in determining cut-off score; hence there is a need for standardization because every teacher sets standardization procedures in its own way. An important fact to remember is that the choice of standard setting method has both psychometric and policy implications. One approach focused, therefore, was the use of criterion referenced setting standards across the higher education and a student who fits to the descriptions of each course or program is evaluated properly so that employers would develop confidence in the recruitment process for the performances are comparable (Cizek, 2001).

To sum-up, there can be different views for explaining the same concept by different individuals. Likewise, this is what is happening for these two concepts (norm and criterion). The core objective of this paper is distinguishing the different methods, leaving this issue (which one is better to be employed) for further investigation. But, there is no option other than to adopt and adapt to the criterion referenced standard setting. Once the need to set a performance standard has been established, the following question arises: What is the best method to use to set performance standards? No one standard setting method is agreed upon as the best. In truth, the best method is the one which best fits the characteristics of both the assessment on which standards are being set and the population to whom the standards will be applied (Morgan and Perie, 2004). Because it is possible that different standard setting methods may result in different recommended cut-scores, it is essential that careful thought goes into the decision of which standard setting method to use? Part of this thought process should include consideration of the arguments defending the validity of the use of a standard setting method for the assessment for which it will be used. Additional thought should be given to the type of evidence or documentation which should be collected and maintained

during the standard setting process. One of the purposes of these standardized tests is to increase "accountability" among educators and students. As such, students are expected to meet some of standard proficiency that the tests are designed to assess. Ideally, this standard will be the embodiment of the learning objectives. The standard should present "mastery" of learning objectives, some level of basic proficiency to move on to the next level (Ricker, 2006). Therefore, it is a basic requirement for the universities and colleges to define the level of proficiency for an "A", "B", "C", "D", and "F" student performances for all courses offered by the departments. Besides it should be publicized to all interesting stakeholders mainly to the students.

It is therefore recommended that in a university situation where the judges are from its community within the faculty, at least, should use the Extended Angoff Method. The reason for selecting this method is because it is simple to understand for intellectuals in the universities and colleges. Besides it can be used for classroom situation where there are few students per class. In this respect, two-phase steps (Morgan and Perie, 2004; Hambleton 1978), for standard setting for classroom can be used with some modifications in the context of higher education system in Pakistan, which can be seen in Appendix A.

References

Berk R. A. (1986) Consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56 (1), 137-172. (Available online at: http://links.jstor.org, retrieved on 10-1-2007).

Cherry, T. L. & Ellis, L.V. (2005) Does rank-order grading improve student performance? Evidence from a classroom experiment. *International Review of Economics Education,* 4, 9-19 (Available online at:http://econltsn.ilrt.bris.ac.uk/iree/i4/cherry.htm, retrieved on 15-1-2007).

Cizek, G. J. (2001) *Setting performance standards: concepts, methods and perspectives*, Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Garcia-Quintana, R. A. & Mappus, M. L. (1980) Using norm-referenced data to set standards for a minimum competency program in the state of South Carolina: a feasibility study. *Educational Evaluation and Policy Analysis,* 2 (2), 47-52.

Hambleton, R. K (1979) *On the use of cutoff scores with criterion referenced tests in Instructional setting* (Available online at: http://www.jstor.org/view/00220655/ap050056/05a00060/0, retrieved on 10-1-2007).

Hambleton R. K. & Plake B.S. (1998) Using an extended angoff to set standards on complex performance assessments. *Applied Measurement in Education,* 8(1), pp 41-55 (Available online at: http://www.leaonline.com/doi/pdfplus/10.1207/s15324818ame0801_4, retrieved on 23-1-2007.

Kane M. (1994) Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461 (Available online at: http://www.jstor.org/cgi-bin/jstor, retrieved on 20-1-2007).

Linn R. L. (1978) Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement,* 15(4), 301-308 (Available online at: http://links.jstor.org/sici?sici=00220655(197824)15%3A4%3C301%3CO%3B2-Y, retrieved on 23-1-2007).

Measurement Research Associates (2004). *Criterion referenced performance standard setting* (Available online at: http://www.measurementresearch.com, retrieved on 13-1-2007).

Mills, C. N. (1983) *Comparison of three methods of establishing cut-off scores on criterion referenced tests* (Available online at: http://www.jstor.org/view/00220655/ap050075/05a00080/0, retrieved on 11-1-2007).

Morgan, D. L. & Perie M. (2004) *Setting standards in education: choosing the best method for your assessment and population*. Unpublished paper Educational Testing Service (Available online at: http://cramer.myweb.uga.edu/6600/MorganPerieSettingStandardsinEducation.pdf, retrieved on 23-1-2007).

Plake B. S., Impra J. C. & Irwin P. M. (2000) *Consistency of angoff-based predictions of item performance: evidence of technical quality of results from the angoff standard setting method* (Available online at: http://www.jstor.org/view/00220655/ap050144/05a00050/0, retrieved on 10-1-2007).

Ricker, K., L. (2006). *Setting Cuts ores: Critical review of angoff and modified-angoff method* (Available online at: http://www.education.ualberta.ca/educ/psych, retrieved on 15-1-2007).

Saeed, M. (2002) *Assessment and examinations – a manual for teachers and teacher educators. Lahore:* Directorate of Staff Development, Punjab.

Skakun, E. N. & Kling, S. (1980) Comparability of methods for setting standards. *Journal of Educational Measurement,* 17, 229-235.

Stephenson, A., Emore, P. B. & Evans J. A. (2000) *Methods plainly speaking standard setting techniques: an application for counseling programs,* _____.

Vos, H. (2004) Methods of stting standards for complex performance-based assessments. *Engineering Education and Lifelong Learning*, 14(1-2), 111-120.

Zeiky, M. J. (2001) So much has changed. How the setting of cut-scores has evolved since the 1980s. In G. H. Cizek (ed.). *Setting performance standards: concepts, methods and perspectives* (pp. 19-52), Mahwah, NJ: Lawrence Erlbaum Associates.

Table 1 Measurement of performance level against scores

Score	Percentage	Performance level
95-100	90-100%	A
85-95	80-90%	B
75-85	70-80%	C
65-75	60-70%	D
Less than 65	< 60%	F

Table 2 Comparison of the Two Approaches: CRSS and NRSS

LG	**Absolute Scale, standard-based, or Criterion-referenced**	**Relative Scale, Norm-referenced**
A	• Firm command of knowledge domain • High level of skill development • Exceptional preparation for later learning	Far above class average
B	• Command of knowledge beyond minimum • Advanced development of most skills • Has prerequisites for later learning	Above class average
C	• Command of only the basic concepts of knowledge • Demonstrated ability to use basic skills • Lacks a few prerequisites for later learning	At the class average
D	• Lacks knowledge of some fundamental ideas • Some important skills not attained • Deficient in many of the prerequisites for later learning	Below class average
F	• Most of the basic concepts and principles not learned • Most essential skills cannot be demonstrated • Lacks most prerequisites needed for later learning	Far below class average

Source: Zerihun, Z. (2006) Unpublished material, _____.

Table 3 Sample task analysis for Angoff standard setting method

No.	Major standard setting Tasks	Sample knowledge and skill requirements	Sample selection factors	Sample training activities
1	Acquire understand of the context of the standard setting activity and the environment to which the standard will be applied.	• Purpose of the exam • Test specifications and test development • Rational for and consequences of standard setting	• Ability to recognize benefits and limitations of testing • Ability to appreciate consequences of applying a standard • Knowledge of instructional environment	▪ Compare and contrast purpose of tests to other possible purposes. ▪ Explain test development and item writing procedures. ▪ Discuss rational for standard setting.
2	Develop definition of borderline examinee performance	• Characteristics of examinee population • Education and training experiences of examinee population • Examination of performance data (item performance and examinee performance)	• Experience or contact with the population of interest. • Knowledge of levels of proficiency in examinee population	▪ Describe cognitive characteristics of examinee. ▪ Evaluate levels of examinee proficiency on the exam and criterion of interest ▪ Review educational preparation of examinees ▪ Present charts depicting exam statistics and discuss varying levels of proficiency
3	▪ Estimate minimum passing levels (MPLs) for each item. a. read each item and evaluate the correct answer. b. Evaluate the relative difficulty	• Detailed knowledge of the domain being assessed. • Item characteristics that influence difficulty • Examinee characteristics that influence item difficulty • Basic understanding of probability	- Ability to read at the level required by the exam - Knowledge of subject matter. - Analytical skills (written comprehension; reasoning; speed of closure; problem sensitivity). - Number facility and related skills • Ability to	- Reference need to read every option to consider item difficulty. - Practice estimating item difficulty with feedback and discussion. - Explain fallibility of test items as measures of the construct. - Present concept of measurement error associated with individual items. - Demonstrate factors that influence item difficulty (factors

	c. Estimate the proportion of borderline examinee that will provide a correct response. d. Repeat step 3 for each item on the test		concentrate for long periods of time; persistence	related to test content, item format, and linguistics). - Distinguish between "would" and "should." - State the impact of number on probability of guessing correctly. - Propose pacing and related strategies.

Cizek, G. J. (2001) *Setting Performance Standards: Concepts, Methods and Perspectives.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Appendix A

Two-phase steps for developing and validating criterion referenced testing in classroom standards setting

Steps in Phase-I

- Preparation of domain specification for all courses offered in the university.
- preparation of criterion-referenced test specification
- writing test items
- Editing test items
- Determining the content validity by using content specialists and use of item analysis data.
- Further editing if need be.
- Test assembly: determination of number of test items/domain, test item selection, preparation of directions and sample questions, lay out and test booklet preparation, preparation of scoring keys and preparation of answer sheets.
- Selection of standard setting methods and deciding cut-off scores accordingly
- Ongoing collection of reliability, validity and norms information (Hambleton, 1978).

Phase-II Steps

1. Meet with department of pedagogies of education to gain knowledge about the assessment and the goals of the standard setting process.
2. Choose one of the standard-setting methods. (optional if done in phase one)
3. Choose a panel in the faculty.
4. *Write performance level descriptors.*
5. *Train the panelists to use the method (including practice in providing ratings).*
6. *Train the panelists on the content standards and test items.*
7. *Compile item ratings or holistic judgments from the panelists that can be used to calculate cut score(s).*
8. *Conduct panel discussions regarding the judgments and resulting cut score(s) in large and/or small groups.*
9. Present consequences or impact data to the panel.
10. Conduct a panelist evaluation of the process and their level of confidence in the resulting standards (Morgan and Perie, 2004).

Correspondence

Muhammad Naveed Khalid, Ph.D Scholar
Email: Muhammad_naveed_786@hotmail.com

Muhammad Saeed, Ph.D
Email: drsaeed61@hotmail.com