Educational Sciences: Theory & Practice - 13(4) • 2169-2175
©2013 Educational Consultancy and Research Center
www.edam.com.tr/estp
DOI: 10.12738/estp.2013.4.1829

A Comparison of Bookmark and Angoff Standard Setting Methods*

Sevda ÇETİN[a]
Hacettepe University

Selahattin GELBAL[b]
Hacettepe University

Abstract

In this research, the cut score of a foundation university was re-calculated with bookmark method and with Angoff method, each of which is a standard setting method; and the cut scores found were compared with the current proficiency score. Thus, the final cut score was found to be 27.87 with the cooperative work of 17 experts through the Angoff method. The cut scores derived by calculations using the bookmark method were found as 19.242 for 1 PLM and RP50, as 25.247 for 1 PLM and RP67, as 18.897 for 2 PLM and RP50, and as 25.102 for 2 PLM and RP67. Correlation coefficients are examined between probabilities of expert answers and real item difficulties to see relationship level between experts' determining probability of right answers and the real difficulty of the items. The correlation coefficient between experts' determining probability of right answers and the real difficulty of the items is determined as 0.60. It is find that there is a significant difference between the percent of students whose score is more than qualifying score determined by the foundation university which is 35 point and the percent of students whose score is more than cut scores determined by Bookmark Method for RP50, RP67 and Angoff Method; but it is also find out that there is no significant difference between the percent of students whose score is more than cut scores determined by the Bookmark Method with RP50 and RP67; and Bookmark method and Angoff method.

Key Words

Angoff Method, Bookmark Method, Cut Score, Response Probability, Standard Setting.

The purpose of educational standards, which have been developed since the 1980s is to set common targets for student performances (Airasian & Russell, 2008). Pass/Fail or performance level assessments that made based on cut scores, does not only effect students' individual academic achievements but it also effect the school, state, country achievement and whole education system. PISA, PIRLS and TIMMS tests are effective to compare the countries' education levels. Many countries abroad determine their own state and school based academic qualification standards (Kubiszyn & Borich, 2007). Setting standards in education, on the other hand, is defined as the process of determining one or more cut scores for a test. The role of cutoff scores is to form performance categories by dividing the test scores scale into two or more areas, and thus to classify the individuals. Therefore, cut scores (which have recently been called performance standards, mostly) have become more necessary and more important. According to Cizek (2001), however, standard setting is the whole of the processes of defined systematic rules pursued in determining the scores distinguishing the two or more degrees of performance.

[*] This study includes a part of a Ph.D. thesis by Sevda ÇETİN.

[a] **Sevda ÇETİN, Ph.D.,** is currently a doctor of Measurement and Evaluation in Education. Her research interests are standard setting methods, cut scores, educational statistics and scale development. *Correspondence:* Dr. Sevda ÇETİN, Hacettepe University, Educational Sciences Department, Beytepe, Ankara, Turkey. Email: tsevda@hacettepe.edu.tr Phone: +90 312 297 8550.

[b] Selahattin GELBAL, Ph.D., is currently a professor of Measurement and Evaluation in Education. Contact: Hacettepe University, Educational Sciences Department, Beytepe, Ankara, Turkey. Email: gelbal@hacettepe.edu.tr.

Depending on individual differences, students learn at differing levels at the end of a learning process. While some of them learn all what is taught, some learn less, and some fail to acquire the targeted gains.

Since students' levels of learning are different, performance definitions of different dimensions and different levels should be made. Performance levels are not dependent on the methods of determining cut scores; hence, the levels can also be determined without determining a method.

In the process of determining the levels of performance, firstly the number of categories is established, and general definitions as to what each category means are made. Because dividing into more than three or four categories would make it difficult to distinguish between the differences of levels, it is not desirable to divide into more categories (Zieky & Perie, 2004). By means of performance levels, students' knowledge, skills and abilities in a certain field that can be displayed at a certain level are described in details; in other words, the requirements for a student to reach that level of performance are described.

The process of transforming the performance standard distinguishing people according to their performance levels into figures in the table of test scores can be called as the process of standard setting (Hambleton, 2001). Pursuing certain stages in the process of standard setting, the ultimate result (passing grade or proficiency score) is reached.

There are many methods in standard setting. Hambleton (2001) and Reckase (2006) list some points for determining which method to use as follows:

- The structure of the test items should be observed. Angoff method is very common in use of multiple choice items whereas Bookmark method is more convenient for the tests that have constructed response items and performance evaluation tests.
- Tests which are low reliable should not be used in standard setting process.
- Time that available to set the standard is important. In some methods standard setting process is longer than the other methods.
- Prior experience with a method is important. If researcher has a prior experience with a method, in second experience it may reduce the need for field-testing which can be costly and time consuming.
- Perceptions and evidence about validity of the method is important. For example some researchers would avoid the Angoff method because of concerns about its validity, other researchers have been critical of the contrasting groups method.

As it seen standard setting is very important in test development process. The process of standard setting, in addition being a methodological process, yields much more useful and defendable results when it involves policy makers, test developers and measurement experts (Bejar, 2008).

Several methods of standard setting were introduced in standard setting work conducted, apart from the above mentioned ones. Each method has advantages as well as disadvantages. However, the implementers are undecided about which method to use, when one is more disadvantaged or has more drawbacks. In Turkey also, it is observed that many educators lack knowledge on which method would be more appropriate for which students and for which situations.

Angoff Method

The method recommended by William H. Angoff in 1971 can be used with tests which are not multiple choices in form as well as with multiple-choice tests. The cut score in the Angoff method is composed of predicted values assigned by experts to each question. The alternatives of the items are not evaluated separately in the method, but the item is considered as a whole. In other words, experts predict the response probability of students who are at the border of performance level determined for each question.

Bookmark Method

The method was suggested by Lewis, Mitzel and Green in 1996. Researchers desiring to remove the inadequacies of the Angoff method recommended the method so as to use it in tests containing multiple choice and structured answers, to reduce the work load of experts and thus to facilitate their decision-making, to combine expert decisions with measurement models in determining the cut scores, and to combine the test content with the definitions of performance level (Mitzel, Lewis, Patz, & Green, 2001).

Bookmark method was based on using the Item Response Theory (IRT) and mapping the items. Items are ordered according to the place they occupy in the scale. Their place in the scale is determined according to item difficulty (p). The ordering is from the easiest item to the most difficult item. The reason for calling the method as bookmark is that experts state their decisions with markings in guides where the items are ordered from the easiest to the most difficult. The guides are called ordered item booklets.

If the test is composed of both multiple choice questions and questions requiring structured answers, each question requiring structured answers can appear in the ordered item booklet several times (and the answers of those questions are, partly correct: 1 point, mostly correct: 2 points, and completely correct: 3 points according to the scores).

Bookmark method has often been used recently for several reasons. Firstly, it may be used in mixed item formats- in tests containing both multiple choice questions and questions requiring structured answers. Secondly, the method reduces experts' workload considerably. For instance, if four performance categories are to be distinguished for a 50-item test, then an expert is expected to give 150 (50 items X 3 cut scores) probability values in the Angoff method whereas in the Bookmark method the first cut score is determined in the same ordered item booklet, and the other cut scores are determined by analyzing the other items respectively; and therefore the expert does not have to analyze the same test items again and again. Thirdly, the method is relatively simpler for experts because calculations which are mathematically more complex are completed before the process of standard setting. And finally, since it is an IRT-based method, it also accommodates the advantages of IRT in the psychometric perspective (Cizek & Bunch, 2007). Despite all these advantages, experts can sometimes experience disagreements in terms of ordering the items (Plake, Impara, Buckendahl, & Ferdous, 2005). While they may believe that the order of some items should be changed in the booklet, they may also determine cut scores in differing places (Skaggs & Tessema, 2001). The task undertaken by experts in this method is very different from the one undertaken in the Angoff method or in other test-centered methods. Experts have to make a decision on probability while categorising the items as adequate and inadequate, and thus determining a cut score. Response probability (RP) is the probability of a person of a certain ability level to reply correctly to an item (Huyhn, 2000; Kostald, 2001), and in this item characteristic curve it is equivalent to the probability of responding correctly to item *i* of a person at the θ_k ability level; and mostly the RP is regarded as 0.67 or 0.50 (Huyhn, 2006). Yet, the values between 0.50 and 0.80 are also used (Berberoğlu, 2009; Huyhn, 2006; Karantoris & Sireci, 2006; Zwick, Şentürk, Wang, & Loomis, 2001). Using the response probability shows that the students who are found proficient according to the cut score will answer the items in the front order of the ordered items (the questions in front of the marked question) correctly at the rate of PR (for example 0.67) and that they will answer the items at the back of the ordered items (the questions at the back of the marked question) correctly at a lower rate than the RP (for example 0.67) (Mitzel et al,. 2001; Wyse, 2011).

When the item difficulty (in logits) and the response probability are given, the level of ability required for achievement probability equivalent to the RP can be determined.

This level of ability is the bookmark difficulty location (BDL). Even though the measurement is called a difficulty location, it is actually a measurement of ability (or rather, it is a measure of ability in which ability and difficulty are measured in the same scale).

A difficulty value is calculated for each item and they are ordered in the ordered item booklet from the lowest BDL value (the easiest item) to the highest BDL value (the most difficult item) according to the value of difficulty. Beretvas (2004) calculated the difficulty value BDL for 1-parameter logistic model and 3-parameter logistic model (the 2-parameter logistic model in which chance parameter is regarded as zero) as in the following:

BDL calculations for 1 Parameter Logistic Model

The value of q when P(X = 1|q) = 2/3 needs to be calculated for the 1 Parameter Logistic Model with Response Probability RP=2/3

$$P(X = 1|\theta) = \left(\frac{\exp(\theta - \hat{b})}{1+\exp(\theta - \hat{b})} \right)$$

θ value,

$$\ln(2) = (\theta - \hat{b})$$

$$\theta_{1PL} = \ln(2) + \hat{b}$$

$$\theta_{1PL} = \ln\left(\frac{x}{1-x}\right)+\hat{b}$$

BDL calculations for 3 Parameter Logistic Model

RP=2/3 and The value of p when $P(X = 1|q) = 2/3$ needs to be calculated for the 3 Parameter Logistic Model with Response Probability RP=2/3

$$\left(\frac{1}{1+e^{[-D\hat{a}(\theta-\hat{b})]}}\right)$$

p value,

$$\ln(2) = D\hat{a}(\theta-\hat{b})$$

$$\theta_{3PL}=\left(\frac{1}{D\hat{a}}\right)\ln(2)+\hat{b}$$

$$\theta_{3PL}=\left(\frac{1}{D\hat{a}}\right)\ln\left(\frac{x}{1-x}\right)\hat{b}$$

The Purpose of the Research

This research aims at comparing the cut scores found via Angoff and Bookmark methods- the major methods of standard setting introduced so far- with a passing grade available.

Problem Statement and Sub-problems

Do the cut scores calculated with the Bookmark and the Angoff standard setting methods differ from the current passing grade determined by a foundation University for the English proficiency exam?

The answers are sought in this study to the following questions:

- Is the proficiency score for the University preparatory class English proficiency exam calculated through the Angoff method different from the current proficiency score?
- Is the proficiency score for the University preparatory class English proficiency exam calculated through the Bookmark method different from the current proficiency score?
- What is the level of correlation between the item response probabilities determined by experts in the Angoff method and the real item difficulty calculated from the University prep class English proficiency exam scores?
- Do the percentages of students receiving scores above the proficiency score (the cut score) calculated in the Angoff and the Bookmark methods for the English proficiency exam and the percentages of students receiving scores above the current proficiency score differ?

Method

Type of Research

This research attempts at revealing the advantages and restrictions of the Angoff method and the Bookmark method compared to each other, and at determining the cut score in the mentioned methods. Due to the fact that identification of the properties of the Angoff and the Bookmark methods is related to the identification of a state, this is a descriptive study. On the other hand, because a comparison is made, it is also a basic research study.

The Study Group

The study group was composed of the 564 students who had taken the university English proficiency exam in the fall semester in 2009. The cut scores were determined in cooperation with 17 experts in English language. The language experts were the English language instructors working in the Preparatory schools of various Universities.

The Tool of Data Collection

A proficiency test which had been administered by the preparatory school of a foundation University and which had a cut score was used as the tool of data collection. The 60-item language and vocabulary part of a 125-item multiple choice test was used for our purposes. To make the set of data congruous with the model, 5 items were removed, and the analyses were conducted with 55 items. Thus the exam which 564 students took included 55 items. The test with an average of 28.88 was observed to be moderately difficult (p=0.52). The test, with reliability (a = 0.94), may be said to be discriminative (($\bar{r}$=0.66) also due to the kurtosis value of 0.094 it could be said that there was a distribution a bit more sharp pointed than the normal. Because the coefficient of skewness fell outside the ±1 border, the test observed to be skewed to the left.

Data Analysis

The analyses of the sub-problems were performed in the following steps:

1. The proficiency score for the University prep class English proficiency test (cut score) was determined in the Angoff method in cooperation with 17 experts in English language. The experts

were the English language instructors teaching in preparatory Schools of various Universities. The experts were asked for their opinions on what percentage of students would be able to answer each item correctly by considering the students at the A2 level border (this is the basic English proficiency level determined by the European council for common language framework), that is to say the students at A1 and A2 levels.

2. Firstly, in the process of establishing the cut score in the Angoff method, whether or not there was a compatibility between experts in scoring was found with Kendall's coefficient of W concordance. The proficiency score for the University prep class English proficiency test (cut score) was determined in the Bookmark method in cooperation with 17 experts in English language. In order to be able to determine the cut score in the bookmark method, first the item parameters according to 1PL and 2PL model were predicted for each item in the test; and then they were sequenced according to the item difficulty parameters predicted according to 1-parameter logistic model (b) and 2-parameter logistic model (b), and the serial item guides were formed. While doing this ordering, the response probabilities were found for both models as (RP) 0.50 and 0.67; and thus the bookmark difficulty locations were determined, and the ordering was done considering them.

Prior to determining the cut score through the Bookmark method, the item parameters according to the 1-parameter logistic model and the 2-parameter logistic model were predicted by using the BILOG programme. Based on the item parameters determined, the response probabilities (RP), and different models of IRT, the Bookmark Difficulty Locations (BDL) were calculated.

3. In order to see the levels of correlation between item response probabilities found by the experts through the Angoff method and the one calculated with the real data, the correlation coefficients between experts' item response probabilities and real item difficulty were analysed, and the level of consistency was checked.

4. Whether or not the percentages of students receiving scores above the proficiency score (the cut score) calculated in the Angoff and the Bookmark methods for the English proficiency exam and the percentages of students receiving scores above the current proficiency score differ was examined comparatively after calculating the student percentages. In order to do this, the test was conducted for the difference between the two dependent percentages, and the significance was tested through the z test. The level of significance was regarded as 0.05 in the test process.

Results

The findings with regard to the sub-problems of the research are as in what follows:

Findings Concerning the First Sub-problem

In the Angoff method, the scores obtained through the predictions of 17 experts as well as the individual cut scores of each expert and the final cut score which is the average of all these scores were calculated. The experts' individual cut scores were between 17.55 and 37.90, but on calculating the average, the final cutoff score was found as 27.83.

Findings Concerning the Second Sub-problem

In consequence of the calculations, the cut score for 1 PLM and RP50 value was found to be 19.242 whereas the cut score was found as 25.247 for 1 PLM and RP 67, as 18.897 for 2 PLM and RP50, and as 25.102 for 2 PLM and RP67.

Findings Concerning the Third Sub-problem

It was observed that the experts found 0.21 as the lowest and 0.87 as the highest average in relation to the percentages of students' answering the test items correctly. On examining the item difficulty calculated from the test scores, it was remarkable that the figures were very close to the ones estimated by the experts. The difficulty index for the most difficult item was 0.20 while it was 0.87 for the easiest item. Whereas the average rate of answering the test correctly was 0.50 according to the experts, the item difficulty average calculated from the test scores was found as 0.52. As to standard deviation and variances, it may be said that the experts' decisions and the test structure were quite similar. The correlations between experts' average of item answering probabilities and the real item difficulty was found to be 0.60. This value of correlation, which was at a moderate level, was found to be significant at the 0.01 significance level.

Findings Concerning the Fourth Sub-problem

It was found that there was a difference between the percentage of students receiving scores above 35, the cut score established by the University and the percentage of students receiving scores above the two cut scores determined according to the Bookmark method (that is to say, 19 for RP50 and 25 for RP67) and the percentage of students receiving above 27, the cut score determined according to the Angoff method; but that there were not significant differences between the percentages of students receiving scores above the cut score in terms of response probabilities according to RP50 and RP67 in the Bookmark method and in the Angoff method.

Discussion and Recommendations

The effects of the Angoff method and the Bookmark method, two methods of standard setting, on passing scores were analysed in this study and a comparison was made with proficiency scores available, and the findings are discussed below in the order of introducing the research problems.

In relation to the first sub-problem of the research, the fact that the university administration determined a passing score for A2 level students rather than students at the A1 and A2 border while determining the cut score may be the cause of the difference between the current proficiency score and proficiency score calculated in the Angoff method. Thus, the university-determined proficiency score appears to be higher than it should be. Besides, in an interview with University administration during the research, they stated that one who can answer 60% of this test correctly can pass the exam, and they were observed to reduce the cut score into 33.

In the second sub-problem, the cut scores determined for 1 PLM and 2 PLM and for RP67 in the Bookmark method were found to be higher than those determined for RP50. These findings are in parallel with the ones obtained by Mueller, Schneider, and Egan (2008). This could be ascribed, as is pointed out by Hambleton and Pitomiak (2006), to the fact that when RP50 instead of RP67 is given to the experts they will progress further to the later questions in the serial item guide, and thus it will cause the higher RP value to determine a higher cut score. This is a finding which is supported by Gembe Tshering (2011). Tshering determined a cut score at the level of RP50 and RP67 for one of CITO's examinations for which the passing score was 28. In consequence, the cutoff score was established as 14.9 for RP50 and 23.10 for RP67. This finding might have stemmed from fact that the University administration considered A2 level students rather than students at the A1 and A2 border in determining the cut score, as in the first sub-problem of the research.

In relation to the findings concerning the third sub-problem, it was found that the experts' decisions and the test structure were quite similar, and that there was a moderate level correlation between experts' average of item answering probabilities and the real item difficulty. Brandon (2004) contends that unless there is a high level correlation between experts' prediction averages and the real item difficulty, the predictions are invalid in the Angoff method. Although Hambleton (2001) also suggests that there should be high correlations between the average of expert predictions and item difficulty, Reckase (2000) states that a moderate level correlation is sufficient for this. Impara and Plake (1998) found the correlation between the average estimated item difficulty and the real item difficulty as 0.78; and claimed that this moderate level correlation was adequate for considering experts' predictions valid.

That the significant coefficient revealed by this sub-problem was at the moderate level may be the result of the fact that the experts predicted the response probability for difficult items higher than it should be while they predicted the probability for easy items lower than it should be (Clauser et al., 2009).

It may be said that the difference between the percentages of students receiving scores higher than those determined in the Angoff and the Bookmark methods and the current score stems from the fact that the cut score determined by the administration was determined by considering students at the A2 level instead of a method of standard setting and that the cut scores calculated in the Angoff and the Bookmark methods had been calculated by considering students at the border of A2 level, with the predictions of the experts.

References/Kaynakça

Airasian, P. W., Russel, M. R. (2008). *Classroom assessment. Concepts and applicatioans*. NW: McGraw-Hill Higher Education.

Angoff, W. H. (1971). *Scales, norms, and equivalent scores.* In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington D.C: American Council on Education.

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

Berberoğlu, G., (2009). Madde haritalama yöntemi ve Cito Türkiye öğrenci izleme sistemi (ÖİS) uygulamalarında yeterlik düzeylerinin belirlenmesi. *Cito Eğitim: Kuram ve Uygulama*, *3*, 14-24

Beretvas, N. S. (2004). Comparison of bookmark difficulty locations under different item response models. *Applied Psychological Measurement*, *28*(1), 25-47.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17*(1), 59-88.

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

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

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education*, *22*, 1-21.

Hambleton R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

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* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

Huyhn, H. (2000, April). *On item mappings and statistical rules for selecting binary items for criterion-referenced interpretation and Bookmark Standard setting*. Paper presented at the Annual Metting of National Council on Measurement in Education, New Orleans, LA.

Huyhn, H. (2006). A clarification on the response probability criterion RP 67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Impara J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*(1), 69-81.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark Standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4-12.

Kolstad, A. (2001*). Literacy levels and the 80 percent response probability convention*. In I. Kirsch, K. Yamamoto, N. Norris, D. Rock, A. Jungeblut, P. O'Reilly, … H. Goksel (Eds.), *Technical report and data file user's manual for the 1992 National Adult Literacy Survey* (pp. 348-370). Washington, DC: U.S. Dept. of Education, National Center for Education Statistics.

Kubiszyn, T., & Borich, G. (2007). *Educational testing and measurement: Classroom application and practice* (8th ed.). Hoboken, NJ: Wiley and John Sons Inc.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark Procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum Associates.

Mueller, C. D., Schneider, M. C., & Egan, K. (2008, March). *Response probability criterion and subgroup performance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York City.

Plake, B. S., Impara, C. S, Buckendahl, C. W., & Ferdous, A. A. (2005, April). *Setting multiple performance standards using the Yes/No Method: An alternative item mapping procedure*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Reckase, M. D. (2000). *The evolution of the NAEP achievement level-setting process: A summary of the research and developmental efforts conducted by ACT*. Iowa City, IA: ACT, Inc.

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for Standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4-18.

Skaggs, G., & Tessema, A. (2001, April). *Item disordinality with the Bookmark standard setting procedure.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Tshering, G. (2011). *Setting performance Standard by using bookmark method.* Retrieved from http://info.worldbank.org/etools/docs/library/240256/Day1StandardSetting_GemboBuhtan.pdf

Wyse, A. (2011). The similarity of Bookmark cut scores with different response probability values. *Educational and Psychological Measurement*, *20*(10), 1-23.

Zieky, M., & Perie, M. (2004). *A primer on setting cut scores on tests of educational achievements*. Princeton, NJ: ETS.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15-25.