A New Approach on Standard Setting with Fuzzy Statistical Analysis

Ming-Chuan Hsieh

National Academy for Educational Research

Abstract

Fuzzy bookmark method has some promising features that help resolving the unclear thinking in human logic and recognition, in this paper, the author use fuzzy statistical analysis technique and a novel modified bookmark procedure to find an optimal cutoff score. Empirical study shows that our proposed fuzzy bookmark procedure has the advantages in finding an appropriate standing setting than the conventional one. The comparison results of these two procedures are also presented and discussed.

Keywords：fuzzy logic, bookmark procedure, standard setting, fuzzy mean, membership functions

模糊統計在標準設定上的應用

謝名娟

國家教育研究院

摘要

模糊統計常用來解決人類多元複雜的曖昧與不確定的現象，而本研究則嘗試利用模糊理論來進行學習成就測驗中的標準設定。在測驗領域，書籤標定法為一種最常用的方法。然而，標準設定成員常面臨無法找出一個明確的切點來放置書籤位置，而在兩或三個書籤切點中徘徊猶豫，進而影響到標準設定成員們共識的形成。本研究旨在將模糊理論融入在傳統的書籤標定法中，並用來找出最佳的切斷分數。實徵研究之結果顯示，模糊書籤標定法有其使用上的優點，尤其在成員決定切斷分數時，比傳統的書籤標定法更容易形成共識。

關鍵字：模糊理論、書籤標定法、標準設定、隸屬度

I、Introduction

Standard setting is a very most important task in the measurement community due to changes in standards-based education and public demands for educational accountability. In the early development of standard setting approaches, the most common methods were those developed by Nedelsky (1954), Angoff (1971). Many other methods were constructed based on these three methods. According to Berk (1996), more than 50 standard setting methods have been developed.

The process of setting standards for student achievement can be both challenging and time-consuming. One level (such as minimum pass score) or multiple levels (such as basic, proficient and advance) can be established for different test conditions. For example, for the teacher certification exam, one may want to establish a minimum pass score to identify qualified teachers. However, for an achievement test, one may need multiple standards to allocate students to such categories as basic, proficient or advance (Sireci, Hauger, Wells, Lewis, Delton, & Zenisky, 2007).

Most standard setting methods calculate the mean, medium and mode and use these statistics as inferences for setting cutoff score. For example, experts in the Yes/No Angoff method indicate whether a borderline student can correctly answer each checklist question. If this borderline student can respond the item correctly, write down “yes” on the recording sheet; otherwise, write down “no”. The mean value is then used as the cutoff score. Similarly, in the conventional bookmark procedure, experts were asked to put their bookmarks using a booklet constructed with items ordered from easiest to hardest. Experts must then find the location at which their borderline students would likely get all items correct and likely get all items beyond that point incorrect. Experts placed a bookmark at this point. The mean or median theta values associate with experts’ placement locations are the references of cutoff score. However, human thought can be incomplete and uncertain. For example, an expert may place a bookmark on item 7. However, some experts may not have the same reasoning and may believe that a student’s ability is a continuous variable, and may place the bookmark between item 7 and 8. In this case, one must compute the information using fuzzy logic.

Human thinking can be designed as formal thinking and fuzzy thinking. Formal thinking is based on logic and order, whereas fuzzy thinking is based on overall conditions combined with different elements. When making a decision, describing the logic in human thinking based on the binary logic of formal thinking is always difficult.

Basic statistical parameters, such as mean, medium, and mode, are important in standard setting procedures. Although conventional methods for calculating these statistics must deal with a single answer or a certain range for an answer, this usually cannot reflect incomplete and uncertain thoughts. These processes often ignore the complex and sometimes conflicting human logic and feeling. If people can use a membership function to express the degree of their perceptions of certain ideas, the outcome will be close to their real thought.

The purpose of this work is to demonstrate how to apply fuzzy statistics to the standard setting procedure. Another goal is to compare the efficacy (e.g the easiness of obtaining consensus among panelists) of this fuzzy bookmark procedure with the conventional bookmark procedure. Additionally, the main ideas in fuzzy theory are also illustrated.

II、Research Method

A、Bookmark Method

The conventional bookmark procedure is item-response-theory-based and has been widely implemented by many testing programs (Lewis, Mitzel, & Green, 1996) for the following reasons. (1)This method can be used for mixed format assessments; that is, this method can easily handle an assessment consisting of both constructed responses and multiple choices items. (2)This method is comparatively easy to implement for experts. Although some computational aspects with this method are complex, most work can be completed before the standard setting session starts. (3)Most high stakes assessments are undertaken in accordance with the Item Response Theory model. This method is also based on IRT. Via this procedure, the analyses normally carried out when constructing and equating these tests make IRT-based standard setting procedures a natural extension (Cizek & Bunch 2007).

Moreover, conventional bookmark method is an IRT-based item-mapping procedure. The items are ordered from least difficult to most difficult and compiled in a booklet. As described by Lewis et al. (1996), item order is established by locating this order on an ability scale at the point at which a borderline student would have a 0.67 probability of success. Because experts do not conduct an item-level examination relative to absolute item difficulty, judgment burden is reduced (Lewis et al., 1996).

After the judges reviewed the performance level description and participate in training activities, they conceptualize borderline students they have taught. With the definition of borderline students have a 67% chance of answering an item correctly (Huynh, 2006 ; Williams & Schulz, 2005); experts are given a booklet of items ordered from least difficult to most difficult based on three parameter IRT. The experts then begin working through the booklet and sequentially determine whether borderline students would have a 67% likelihood or better of answering items correctly. At some point in the ordered item booklet (OIB), experts identify when the likelihood that borderline students would answer correctly drops below 67%. The associating page number at that point is recorded in the OIB. After the judges make this decision for the basic category, they continue examining items past the bookmark point to identify the cutoff scores for proficient and advance students.

The sheet on which teachers recorded the page number for each performance level were collected, data were entered and averaged cutoff scores for each level were calculated. The experts were provided with feedback, including graphs showing the distribution of teacher bookmarks and the cumulative percentage of students at each performance level. After viewing the feedback, the experts had a group discussion. The experts exchanged opinions about what makes one item or group of items more difficult than preceding items and, ultimately, placed a bookmark at the point where they believe the difficulty of subsequent items exceeds the ability of an identified group of students. The experts were then asked to make a second bookmark based on how they expected borderline students to perform. They were provided with the same feedback and asked to perform a third run for bookmark placement. The recommended cutoff score is based on the third bookmark placement task and the average theta values for experts.

B、Why we use Fuzzy Bookmark Method?

Since Zadeh (1965) first developed fuzzy set theory, its applications have been extended to conventional statistical inferences and methods in the social sciences. In reality, both fuzzy theory and probability theory deal with problems with uncertainty. Nguyen and Wu (2002), developed many approaches and concepts based on fuzzy theory. Additionally, Lowen (1990), Tseng & Klein (1992), Wu & Sun (1996), Kosko (1993), Guariso, Rizzoli, & Werthner (1992) introduced many important ideas and methods into computations in the social sciences.

When 1 represents "Yes," and 0 represents "No," a clear answer must to identify whether an answer is 1 or 0. However, some answers may be 1 or not 0. This situation can be logical and reasonable. For example, a person can be smart, but not too smart. Based on this logic, fuzzy theory is suited to the real world.

Many issues associated with experimental education policies, especially those related to social benefits, should seek public consensus as a reference when making decisions and evaluation. During standard setting, the assessment task involves cooperation among teachers or educational specialists. Hence, the ability of teachers and educational specialists to express real feelings or thoughts of teacher is crucial in the standard setting meeting.

The following section presents some basic ideas in fuzzy statistics.

C、Fuzzy Mean (Nguyen and Wu, 2002)

Let $\underline{A} = \frac{f_A(x_1)}{x_1} + \frac{f_A(x_2)}{x_2} + ... + \frac{f_A(x_n)}{x_n}$ be the universal set (a discussion domain), $L = \{L_1, L_2, \cdots, L_k\}$ be a set of k discrete variables on U, and $\{FS_i = \frac{m_{i1}}{L_1} + \frac{m_{i2}}{L_2} + ... + \frac{m_{ik}}{L_k}, i = 1,2,..., n\}$ be a sequence of a random fuzzy sample on U; $m_i (\sum_{j=1}^{k} m_{ij} = 1)$ is the memberships with respect to L_j.The fuzzy mean is then defined as

$$E(X) = \frac{\sum_{i=1}^{n} m_{i1}}{n} \frac{\sum_{i=1}^{n} m_{i2}}{n} + ... + \frac{\sum_{i=1}^{n} m_{ik}}{n}.$$

In the fuzzy bookmarking procedure, teachers recorded up to 3 possible bookmark locations for each performance level, and assigned a weight to each location. The overall fuzzy mean was calculated as a final cutoff score.

D、Data Description

A one-day workshop for setting cutoff scores for a Grade 6 English test was held for a panel of 31 experts. To ease the comparison of conventional and fuzzy bookmark method, each panelist is asked to conduct both methods in the workshop. These experts were elementary school teachers (n=24), university professors (n=3), and school principals (n=4). These experts were selected or recommended by a content specialist, represented cities and rural areas in Taiwan. The experts were provided with an overview of the session purpose and their objectives a week before the

workshop. The workshop began with discussions of the purpose scoring tests and provided experts with a clear understanding of test content. A discussion of performance level descriptions (PLDs) followed. Once experts understood the PLDs, they conceptualized the definition of borderline students for each level.

To familiarize experts with item types and the range of item difficulty on the OIB, a short practice activity was conducted. After completing these practice activities, the experts discussed their experiences and completed an evaluation form to assess their understanding of the task and their readiness to initiate the bookmarking procedures.

E、Training Procedures: Bookmark and Fuzzy Bookmark Procedures

The training was divided into two parts. The first part is the illustrate how to implement the conventional bookmark procedure. For the conventional bookmark procedure, experts were divided into 5 groups, and each with 6–7 experts. Each expert received a copy of the OIB. By thinking of the borderline group for a specific performance level, the experts started with the least-difficult item and progressed through the booklet until they reached the point at which they believed borderline students would likely have a 67% chance of getting all items correct up to that point. Students should have a < 67% chance of getting all items correct after that point. The experts were told to place a bookmark at that point. After panelists completely understand the conventional bookmark approach, the facilitator started to illustrate how to implement the fuzzy bookmark procedure. The fuzzy bookmark procedure is quite similar to the conventional approach, the difference is experts were asked to write down up to three likely bookmark locations for each performance level and assign the membership functions. Table 1 shows the recording sheet.

The page numbers for each performance level were collected, data were entered, and averaged cutoff scores for each level were calculated. The experts were provided with feedback, including graphs showing the distribution of each teacher’s bookmark placement and the cumulative percentage of students at each performance level. After viewing this feedback, a group discussion was held. Discussions focused on reasons for bookmark locations and item characteristics as well as the knowledge and skills of borderline students. Each group was informed that variability among teacher bookmarks was expected.

Experts then discussed the impact of the range of cutoff scores. This range included the rounded average (mean or median), and a scatter plot was provided for the bookmarks of each expert. Additionally, the cumulative percentage distribution calculated from actual student performance was also provided.

F、Assessment description

The Grade 6 English test assesses the English-language skills of Taiwanese students was used. The test purpose is to determine the effectiveness of elementary school instruction and curriculum as well as the general performance of Taiwanese students. Roughly 10,000 randomly sampled students take this test annually. A cutoff score for this test was used to classify students into three categories—basic, proficient, and advanced. All items are multiple-choice items with 3 response options. The test has 103 items.

Table 1 **Recording sheet**

Panelist Number:

Direction: Enter your bookmark page number for each performance level in the space below. If each performance level should have more than one possible location, record up to three bookmarks and their weights. Examples are as follows:

Basic		Proficient		Advanced	
Page#	Weight	Page#	Weight	Page#	Weight
10	60%	35	20%	78	30%
11	40%	36	60%	79	70%
		37	20%		

Final Bookmark page No.			
	Basic	Proficient	Advanced
Page #			

Round 1 (same sheet for rounds 2 and 3)

Basic		Proficient		Advanced	
Page#	Weight	Page#	Page#	Weight	Page#

Final Bookmark page No.			
	Basic	Proficient	Advanced
Page #			

III、Results

Experts were provided with bookmark placements before and after receiving actual student performance data. Table 2 presents the cutoff scores for the bookmark procedure in each run. Providing actual performance data between rounds had some influence on the experts' placement of the bookmark. The theta cutoff scores from round 1 increased in round 2 for all three performance levels. However, following the discussion before Round 2, the proficient level and advanced level cutoff points declined. At the start of the sessions, experts were only aware of relative, not absolute, item difficulty during the Round 1. After providing experts with the percentage of students at each performance level, experts adjusted their decisions.

With the proposed fuzzy bookmarking method, most theta cutoff scores were slightly lower than those obtained the conventional bookmarking method. However, the standard deviation for the fuzzy bookmark method was generally lower than that of the conventional bookmark method. This decreased standard deviation produced a narrow range of possible cutoff scores, indicating a high level of inter-expert agreement.

When comparing Table 2 and 3, it shows that the cut scores and standard deviation for the advanced level are identical. It is because the panelists are quite certain about the bookmark location for the advance level. Although the facilitator asked the panelists to write down three possible bookmark locations in the fuzzy bookmark method, panelists still prefer to just given a single bookmark location which they are quite certain about. In this special case which the panelists just give one bookmark location, the resulting cut scores and standard deviation for both fuzzy and conventional bookmark method will be the same.

Table 4-5 and Figure 1-2 list the percentage of students in each performance level obtained using these two standard setting procedures. For example, the greatest variability in percentage of students occurred in Round 1, especially for the basic and proficient levels. With the conventional bookmark procedure, standard deviation is 0.16 for the basic level at round 1; this percentage decreased to 0.14 with the proposed fuzzy bookmarking procedure. With the conventional bookmarking procedure, SD is 0.28 for the proficient at round 1; this percentage decreased to 0.26 with the proposed fuzzy bookmarking procedure. But the differences are smaller at round 2 and 3.

Workshop evaluation data were also collected to assess the level of confidence in expert determinations. Ratings were on a five-point Likert scale, with 5 for "strongly agree," 4 for "somewhat agree" 3 for "neutral," 2 for "disagree," and 1 for "strongly disagree". Analytical results demonstrate that experts had a high degree of confidence in the passing score (4.5 with the conventional bookmarking procedure and 4.25 with the proposed fuzzy bookmarking procedure). The score is slightly higher for the conventional procedure. It is because some panelists have difficulty in giving membership function for the fuzzy method, thus, how to implement fuzzy method properly that widely accepted by participants is something need to careful for future researchers.

Table 2 **Bookmark method cut score means and standard deviation**

Round	Cut theta score (SD)		
	Basic	Proficient	Advanced

Table 2 **Bookmark method cut score means and standard deviation (cont.)**

1	-0.90 (0.16)	-0.34(0.28)	0.21(0.22)
2	-0.94 (0.11)	-0.43(0.23)	0.30(0.17)
3	-0.96 (0.06)	-0.08(0.50)	0.22(0.10)

Table 3 **Fuzzy Bookmark method cut score means and standard deviation**

Round	Cut theta score (SD)		
	Basic	Proficient	Advanced
1	-0.89 (0.14)	-0.32(0.26)	0.21(0.22)
2	-0.94 (0.09)	-0.40(0.26)	0.30(0.17)
3	-0.95 (0.06)	-0.10(0.48)	0.22(0.10)

Table 4 **Percentage of students in each performance level by bookmark procedure**

Level	Round1	Round2	Round3
Below Basic	20	19	19
Basic	16	14	25
Proficient	18	25	11
Advanced	46	42	45

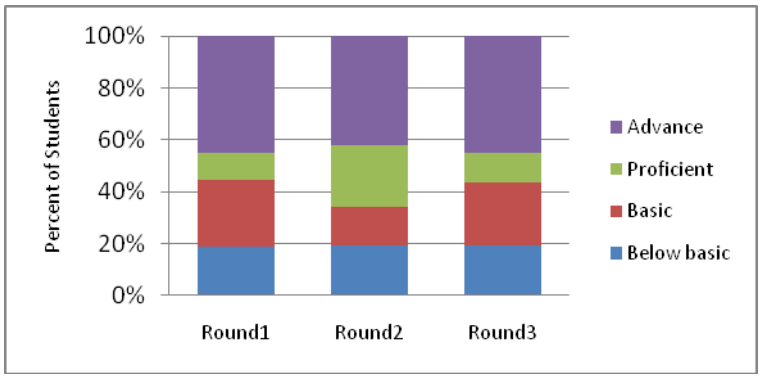

Figure1 **Percentage of students in each performance level by bookmark procedure**

Table 5 **Percentage of students in each performance level by fuzzy bookmark procedure**

Level	Round1	Round2	Round3
Below Basic	19	19	19
Basic	25	15	24
Proficient	11	24	11
Advanced	45	42	45

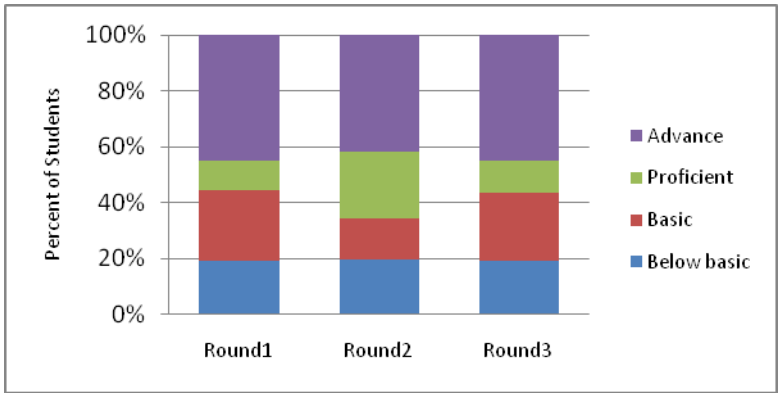

Figure 2 **Percentage of students in each performance level by fuzzy bookmark procedure**

IV、Conclusion and Discussion

One difficulty typically encountered when implementing the conventional bookmarking method is that some items do not follow the difficulty order determined by experts, especially when items are ordered based on their difficulty level using the 3PL item response model. Additionally, the difficulty level is confounded by discrimination and guessing parameters. Experts may have difficulty finding a single location to place a bookmark. However, the fuzzy bookmark procedure allows experts to place multiple bookmarks for each performance level; that is, when experts are uncertain about a single bookmark, they can place multiple bookmarks to express their true feelings. This procedure is in the spirit of fuzzy statistics. Bellman & Zadeh (1970) indicated the argument of fuzzy decision-making is that people's attitudes and feelings should have different degrees of membership, and that fuzzy theory is suitable when analyzing human attitudes and feelings. When applying fuzzy theory to standard setting, one can easily assess the true feelings of experts about their decisions.

Analytical results also demonstrate that the fuzzy bookmarking procedure has some important benefits. At least for the early rounds, the proposed fuzzy bookmarking procedure produced a smaller standard deviation than the conventional bookmarking procedure. As Berk (1996) indicated, reducing inter- expert variability is a positive outcome; thus, the proposed fuzzy bookmark procedure has its advantages. However, how to further utilize the fuzzy statistics in standard setting needs more researches.

Because the application of the fuzzy bookmarking method was limited to an English test, studies examining other subjects or with different applications would prove beneficial. For instance, will similar outcomes be obtained for a certification setting, where experts may be relatively less familiar with the characteristics of examinee populations? Additional applications of the fuzzy bookmarking method are needed, as are examinations of the understanding of experts of standard setting procedures. Other fuzzy statistics, such as fuzzy median and fuzzy mode, should also be utilized in future studies. Finally, additional comparisons with other methods (such

as Angoff method) are needed to understand the characteristics of outcomes.

References

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

Bellman, R. E., & Zadeh, L. A. (1970). Decision-making in a fuzzy environment. *Management Science, 17*(4), 141-164.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9*(3), 215-235.

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

Guariso, G., Rizzoli, A., & Werthner, H. (1992). Identification of model structure via qualitive simulation. *IEEE Trans. on Systems, Man, and cybernetics, 22*(5), 1075-1086.

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

Kosko, B. (1993). *Fuzzy thinking: the new science of fuzzy logic*. New York: Hyperion.

Lewis, D. M., Mitzel, H.C., & Green, D. R. (1996). *Standard setting: A bookmark approach.* Paper presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Boulder, CO.

Lowen, R. (1990). A fuzzy language interpolation theorem. *Fuzzy Sets and Systems, 34*, 33-38.

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

Nguyen, H., & Wu, B. (2002). *Fuzzy Mathematics and Statistical Applications*. Taipei: Hua-Tai Book Company.

Sireci, S. G., Hauger, J., Wells, C. S., Lewis, C., Delton, J., & Zenisky, A. (2007). *Evaluation of the standard setting on the 2005 Grade 12 National Assessment of Educational Progress math test.* Center for Educational Assessment Research Report No. 618. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Tseng, T., & Klein, C. (1992). A new Algorithm for fuzzy multicriteria decision making. *International Journal of Approximate Reasoning, 6*, 45-66.

Williams, N., & Schulz, E.M. (2005). *An investigation of response probability values used in standard setting*. Paper presented at the meeting of the National Council on Measurement in Education. Montreal, Canada.

Wu, B., & Sun, C. (1996). Fuzzy statistics and computation on the lexical semantics. In *11th Pacific Asia Conference on Language, Information and Computation* (pp. 337-346). Seoul: Kyung Hee University.

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control, 8*, 338-353.