Cross Validating Modified Angoff and Bookmark Standard Setting for a Home Inspection Certification

James B. Olsen and Russell Smith, Alpine Testing Solutions
Paper Presented at the Annual Meeting of the National Council on Measurement in Education, New York, March 2008

Abstract: This paper presents a cross validation of a standard setting evaluation for a home inspection certification program. Two performance standard setting approaches, the Modified Angoff and the Bookmark, were implemented with independent subject matter expert panels to specify a cut score region for certification board review and recommend a preliminary cut score. Performance standards were set on a home inspection exam with 120 multiple choice test items. The two standard setting procedures produced very similar results when evaluated using the target standard setting regions, the preliminary cut scores and the standard errors for the judge ratings. The results of the standard setting study are discussed.

ISSUES IN PERFORMANCE STANDARD SETTING

Ronald K. Hambleton notes that the setting of performance standards is "the most controversial problem in educational assessment today." (Hambleton, 1998, p. 103). In the current accountability environments for schools, higher education, accreditation, conformance assessment, certification and licensure setting performance standards remains an issue of high visibility, criticality, and discussion from multiple perspectives. Cizek (2006) states,

> "Thus, agencies responsible for testing programs must recognize that, as long as important decisions are being made, and as long as test performance plays a role in those decisions, it is likely that controversy will remain. At least to some degree, however, the defensibility of those decisions will be maximized by crafting well conceived methods for setting performance standards, implementing those methods faithfully, and gathering sound evidence regarding the validity of the process and the result." (Cizek, 2006, p. 256)

In Michael Kane's (2006) recent discussion for test validation, he notes the crucial importance of standard setting in interpreting the validity of certification and licensure test scores and educational performance levels in the No Child Left Behind accountability programs.

> "The interpretive argument for a licensure test typically involves a semantic interpretation of the professional competence as a broadly defined trait variable and then a decision procedure that implements a policy about the level of competence required for admission to practice…The choice of cutscore is the main issue in defining the decision rule, and the evaluation of the cutscore is the key issue in validating the decision rule.

> Standard setting studies are designed to identify a reasonable cutscore and to provide backing for the choice of cutscore."
>
> "Under the No Child Left Behind (NCLB) Act (NCLB, 2002) student scores are transformed to general achievement levels, intended to reflect different levels of performance…The achievement levels are defined by cutscores on the score scale for the test…Interpretive arguments for NCLB accountability programs would involve an initial semantic interpretation of student performance in terms of individual achievement on the state standards (a trait attribute), a conversion of these scores to achievement levels (basic, proficient, advanced), and the computation of the percentages at each level in each grade for a school (and for a subgroup) followed by a decision about the school." (Kane, 2006, pp. 52-53)

Kane's statements highlight the importance of carefully defining the performance standard and its levels, the selected cutscore(s), and the evaluation of the reasonableness of the cutscore and the threads of evidence supporting the cutscore and the performance level descriptions.

PERFORMANCE STANDARD SETTING METHODS

Various performance standard setting methods have been developed and used in validating scores and interpretations from certification, licensure and educational tests (Berk, 1986; Cizek, 2002; Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006; Jaeger, 1989, 1993; Plake, 2005, 2007. This paper illustrates a cross validation of the modified Angoff and the Bookmark standard setting methods and procedures within the context of a home inspection certification program.

Modified Angoff Method

Cizek and Bunch (2007) state, "…it is certain that the Angoff method (and all of its variations) is the most commonly used method for setting performance standards in contemporary use in licensure and certification contexts." (Cizek and Bunch, 2007, p. 82) Plake (2005) notes, "The most prevelant method for setting cutscores on multiple-choice tests for making pass-fail decisions is the Angoff method."

Research also shows that the modified Angoff is quire pervasive in certification, licensure, and credentialing programs (Meara, Hambleton, and Sireci, 2001; Sireci and Biskin, 1992; Mills and Melican, 1988; Colton and Hecht, 1981). Ricker (2006) provides a critical review of the Angoff and modified Angoff Standard setting methods.

The Modified Angoff method asks subject matter experts (SMEs) to conceptualize either an examinee or group of examinees who are minimally competent (the minimally competent candidate (MCC)) for a given performance level. For each test item SMEs estimates the probability that the MCC will answer the item correctly. SME ratings are made independently. The item level probability levels are summed across SMEs to determine an overall estimated

passing rate across items and judges. Multiple rounds of the ratings are performed interspersed with feedback specifying minimum passing score, high and low judge ratings, empirical examinee performance data, and impact data for expected examinee passing percentages. Minimum passing scores, high and low judge ratings and panel variability are calculated for each judging round. Two to three judging rounds are typically required to reach convergence on the minimum passing level or cut scores.

Bookmark Standard Setting

Cizek and Bunch (2007) note, "The Bookmark procedure has become quite popular for several reasons. First, from a practical perspective, the method can be used for complex, mixed-format assessments, and participants using the method consider selected-response (SR) and constructed-response (CR) items together. As the prevalence of mixed-format examinations continues to increase, it is likely that the Bookmark will become even more widely used and that other innovative approaches for setting performance standards in such contexts will be developed." (Cizek and Bunch, 2007, p. 157,159)

Karantonis and Sireci (2006) state, "The Bookmark method for setting standards on educational tests is currently one of the most popular standard-setting methods. However, research to support the method is scarce." (Karantonis and Sireci, 2006, p. 4)

With the Bookmark procedure the items are presented in an ordered item booklet with one item or constructed response score rubric point per page from the easiest to the most difficult item. SMEs are presented with the ordered item booklet and asked to place their bookmark where they believe the minimally competent candidate would have a specified response probability (50% or 67% of answering the item correctly). They place their preliminary bookmark on the page after to the specified item in the booklet. They are then asked to go beyond that item and determine how the minimally competent candidate would answer the next series of items. The SME places the final Bookmark for that round at the item location they select. The ability level of the item preceding their bookmark and the ability level of the item that is first to exceed their bookmark are used to estimate using the test characteristic curve the number of items expected correct on the test. The specified response probability (50%, 67%) is called the Bookmark response probability. The individual SME Bookmark estimates are often shared with the group either graphically or numerically and discussion occurs regarding the SME Bookmark placements. Impact data from examinee scores can be shared with the SME panelists. There are usually two or three rounds with the SME minimum passing scores reviewed and discussed for each round.

Recent discussions of the modified Angoff and Bookmark standard setting methods are provided in Cizek, Bunch and Koons, 2004; Cizek and Bunch, 2007; Hambleton and Pitoniak,

2006 and Plake, 2005. Lin (2003) provides a review of the strengths and weaknesses of the Bookmark standard setting procedure.

Validity Comparisons of Standard Setting Methods

Many measurement specialists have recommended research comparisons of results from alternative standard setting methods.

In discussing the comparability of standard-setting results, Richard E. Jaeger (1989, 1993) emphasizes that "A large number of empirical studies have addressed the question of whether different standard-setting procedures when applied to the same competency test, provide similar results. Most research has answered this question negatively. Different standard setting procedures generally produce markedly different test standards when applied to the same test, either by the same judges or by randomly parallel samples of judges." (Jaeger, 1989, 1993, p. 497).

> His summary recommendation was that "The results summarized...show that the choice of a standard-setting method is critical. As Hambleton (1980), Koffler (1980), and Shepard (1980, 1984) suggest, it might be prudent to use several methods in any given study and then consider all of the results, together with extrastatistical factors, when determining a final cutoff score."
>
> Hambleton and Pitoniak (2006) state, "Comparative studies of methods are plentiful in the research literature and will not be reviewed here. Often these studies are inconclusive, and do not produce generalizable results for the assessment field, because of many factors: (1) the test may unique in some way, (2) the methods were implemented in ways that may be unique to the particular research studies, and so on. Recent empirical studies may, however be helpful to those interested in reviewing relatively recent comparative studies of methods (e.g., Brandon, 2002, 2004; Buckendahl, Smith, Impara and Plake, 2002; Cizek, 2001; Green, Trimble and Lewis, 2003, Hurtz and Auerbach, 2003)." (Hambleton and Pitoniak, 2006, pp. 450-451).

As a discussant on standard setting papers presented at the 2007 NCME conference, Hambleton (2007) emphasized the need for more research and validation within and across alternative standard setting methods. Plake (2007) highlights thirteen areas for conducting research in standard setting. She notes that, "Although research programs support many of these [standard setting] methods, many of the design and implementation issues that surround the process have not been well researched…With research to support standard setters, we will be

better able to make informed decisions when designing and implementing standard setting studies."

Cizek and Bunch (2007) provide an alternative view recommending caution regarding use of multiple methods of standard setting. "We are aware of only a few contexts in which multiple standard setting methods were used. We are not aware of even a single documented instance in which a systematic, replicable process has been documented for synthesizing the results of the multiple procedures….little progress has been made in research and development of methods for combining the results of multiple standard setting procedures. No methodology currently exists for satisfactorily addressing the challenge that arisen when multiple standard setting procures result in different answer to the standard-setting question." (Cizek and Bunch, p. 329-320)

RESEARCH WITH MULTIPLE STANDARD SETTING METHODS

Poggio, Glassnapp and Eros (1981) compared results from independent samples of teachers that used four different standard setting methods: The Angoff, the Ebel, the Nedelsky and the contrasting groups method. Their study indicated wide variability in the performance standards established by the four different methods at each grade.

Cross, Impara, Frary, and Jaeger (1984) implemented three methods for establishing minimum standards for the National Teacher Examination. Fifteen panelists provided ratings on the mathematic exam and fifteen panelists provided ratings on the elementary teaching exam. They implemented the Angoff, Nedelsky and Jaeger procedures. The Nedesky procedure asked SMEs to identify the distractors that a minimally qualified examinee would identify as incorrect. The results showed that there were substantial differences in the standards that were set across methods and in many cases across test sections. When the results were summarized across sessions and examinations, the percent correct standards that would be set for the Angoff, Jaeger, and Nedelsky methods were respectively 45.37, 60.77 and 29.41. An analysis of these mean scores showed statistically significance differences at the 0.05 level. Previous studies of the Nedelsky procedure showed lower standards than other methods.

Norcini, Lipner, Langdon, and Strecker (1987) compared three variations of the Angoff standard setting method for a gastroenterology specialty examination. Modified Angoff judgments were made before, during or after a SME committee meeting. Six SMEs were involved in the rating The overall Angoff ratings were 59.8 for the before meeting ratings, 63.5 for the during meeting rating and 61.7 for the after meeting ratings. Standard deviations for the ratings were 5.8 for the before meeting rating, 2.4 for the during meeting, and 1.7 for the after meeting rating. Means for the three Angoff modifications showed no statistical differences among variations. The standard deviations of the ratings showed decreases from the before meeting to the during meeting and after meeting conditions.

Koffler (1980) compared results from the Nedelsky and Contrasting Groups procedures for setting standards on statewide minimum competency exams in reading and mathematics for

grades three, six, nine and eleven. The result showed no consistent patterns of agreement or disagreement between the cut scores produced by the Nedelsky and Contrasting Groups methods. At grade nine, the two standard setting methods yielded the same cut score, for grade 11 reading there was a thirty-three point difference, and for grade 11 mathematics there was a thirty-four point difference between the standards set by the different methods. Due to these differences Koffler recommends that "a number of procedures should be used." (Koffler, 1980, p. 177)

Mills (1983) compared result from the Angoff, Contrasting Groups and Borderline Group methods for setting standards from twelve field tested forms for mathematics and language arts for grade two. The Angoff and Contrasting Group standard setting methods provided consistent cut scores for several of the field test forms, however, the same judges made the ratings for the two different methods. The results from the Angoff and Contrasting Groups were more similar and discrepant from the results of the Borderline Group method.

Green, Trimble, and Lewis (2003) compared the Bookmark, Contrasting Groups and Jaeger-Mills (holistic work performance judgment) procedures for setting standards for statewide assessment exams. The results showed that the Bookmark Procedure produced lower cut scores than the other two methods. The Jaeger-Mills procedure yielded cut scores that were higher of the three procedures.

Buckendahl, Smith, Impara, and Plake (2000, 2002) compared the Angoff and Bookmark procedures for setting standards for a grade 7 mathematics examination with 69 test items used with a Midwestern school district. The results showed that the Angoff and Bookmark cut scores were very similar with the Bookmark producing a lower standard deviation. A group of 23 SMEs were used for the ratings. The summary Angoff cut scores were 34.92 for Round 1 and 33.43 for Round two with standard deviations of 7.79 for Round 1 and 10.96 for Round 2. The summary Bookmark cut scores were 33.64 for Round 1 and 33.64 for Round 2 with standard deviations of 11.03 for Round 1 and 8.66 for Round 2. This paper is the most comparable paper to the research presented herein.

Davis, Buckendahl, Chin and Gerrow (2008) also compared the modified Angoff and Bookmark procedures for setting standards for an international licensure program. Thirty-four panelists were divided in to two groups that were counterbalanced and evaluated the test with one group of 17 panelists using the Modified Angoff followed by the Bookmark procedure and a second group of 17 panelists using the Bookmark followed by the Modified Angoff procedure. The results showed that the mean test scores for the different standard setting procedures were within two score points, the median scores were within one score point and the standard errors were within one score point for the different standard setting methods. Also the impact means were within two percentage points and the impact medians were within one half a percentage point for the Bookmark and Modified Angoff standard setting procedures.

Reckase (2006) provides a simulation study to evaluate the results from the Modified Angoff and Bookmark standard setting procedures. Rasch model IRT procedures were used to provide the underlying statistical framework for the research comparisons. The response probability used in this research was 0.67 although additional response probability variations of 0.50 and 0.75 were investigated. With a simulation condition with error free item parameters and perfect judgments, the results showed the recovery of the simulated performance standard was quite poor at the extremes of the proficiency range. This was caused by the lack of items and calibration statistics at the extremes of the proficiency range. The Angoff procedure simulation was able to recover the simulated cut score.

The Bookmark procedure for choosing the placement at the Previous Item (PI) condition always produced a negative bias that was always below the panelists theoretical Intended Cut Score. The Between Item (BI) condition averaged the Rasch ability values for the before Bookmark item and the immediately after Bookmark item. The Between Item condition produced a cut score closer to the Intended Cut Score.

In a simulation with error free item parameters with fallible judgments. The Modified Angoff simulation with modeled judgment error condition produced a reasonable unbiased estimate of the modeled cut score. The Bookmark simulation with a modeled error condition produced a biased estimate of the cut score that was typically 0.5 theta units lower than the Intended Cut Score. This situation resulted with the simulated panelists placing their bookmarks earlier in the ordered item booklet than would be expected by the response probability of 0.67. Reckase recommends use of both the before item Book mark and after item Bookmard be averaged in computing the theta values for the Bookmark placements. Schulz (2006) provides a response to Reckase (2006) research which shows that the Bookmark standard setting and Mapmark variations. He also notes that higher cut scores are specified based on the higher Bookmark response probabilities. Schulz also notes that the Bookmark procedure typically asks panelists to "go beyond" the first item that they identify that has the expected probability value specified by the project. Schulz data showed that Markmark ratings were similar to Modified Agoff ratings.

RESEARCH HYPOTHESES

The researchers hypothesize that the cut scores from the Modified Angoff and Bookmark standard setting procedures will be the same or the score regions of the mean plus and minus one standard error will be overlapping.

METHODS

THE ANGOFF PROCEDURE

The Angoff procedure is a widely accepted methodology for establishing the performance standard cut score for a test. The procedure relies upon the judgment of SMEs who examine the

content of each test item/task and predict the proportion of minimally-qualified candidates that will answer the item correctly. The average of the judges' predictions for a test item becomes the predicted difficulty of the test item. The sum of the predicted item difficulty values for each item averaged across the judges and items on a test form is the recommended cut score.

The expected Angoff item difficulties were derived from the judgments of a total of eight SMEs. The certification body maintains identity and qualifications of the SMEs.

The Angoff performance standard setting was held June 22, 2007 in Minneapolis, MN at the certification body headquarters. The SME meetings were facilitated by one author of this paper. Eight SMEs completed the first and second rounds of the Angoff standard setting process. The Angoff performance standard setting process included:

1. Review description and purpose of the exam
2. Taking the target exam
3. Overview of the test development and Angoff standard setting procedures
4. Discussion of the borderline qualified candidate
5. Angoff Round 1 Ratings (Item Key was provided with the Round 1 ratings)
6. Round 1 Feedback – Delta differences in colleague ratings and discussions of rationales for each rating
7. Angoff Round 2 Ratings
8. Summarization of Results

Angoff Round 1 and Round 2 ratings were done independently by the SMEs. The results were compiled and discussed once judgments were made for each of the 120 multiple choice items.

THE BOOKMARK PROCEDURE

The Bookmark procedure consisted of a psychometrician rank ordering all of the items for the final form of the examination. This ranking is based on each item's difficulty. An ordered item booklet is created where the first item is the easiest, followed by the next easiest, all the way until the end where the last item is the most difficult. After discussing the purpose of the exam and the definition of the minimally competent candidate, the SMEs are instructed to indicate (or "bookmark") how many of the questions (moving from easiest to hardest) would be answered correctly by the minimally qualified candidate. A second description instructs the SMEs to place a bookmark where they believe that the minimally qualified candidate will probably get all of the questions before the bookmark correct and all of the questions after the bookmark incorrect. The bookmark placement, or question count, and then average across the SMEs. The average is then taken to be the recommended cut score.

The Modified Bookmark standard setting was held July 26, 2007. The meeting was conducted virtually using audio-conferencing software. The SMEs were facilitated by another author of this

paper. This was not the same author that facilitated the Modified Angoff standard setting. Five SMEs completed the first and second rounds of the Bookmark standard setting process. The Bookmark performance standard setting process included the same steps as the Angoff process except that the Round 1 and Round 2 ratings consisted of Bookmark placements rather than item level judgments. The Bookmark Round 1 and Round 2 ratings were done independently by the SMEs. The results were compiled and discussed for the 120 multiple choice items.

RESULTS

Following are summary results for the two standard setting procedures presented in Table 1. The performance standard region for the modified Angoff and the Bookmark overlap for 6 out of 17 score points possible within the target standard region. The preliminary cut scores for the modified Angoff and Bookmark procedures are separated by only 4 score points which is just slightly larger than one standard error out of a total of 120 test items. The modified Angoff procedure showed slightly higher mean, median and standard error judge ratings than the Bookmark procedure. Within the Angoff procedure there was a reduction in the standard region, a reduction in the standard errors, a decrease in the estimated standard error of measurement, and an increase in the reliability and internal consistency of ratings. The Bookmark procedure showed a decrease in the standard region, a large reduction in the standard error for judges; however there was remarkable stability for the SME mean and median rating. Mean and median ratings were rounded to the nearest whole score value.

Figure 1 provides a bar graph summary of the 8 judges Modified Angoff ratings. Figure 2 presents a bar graph summary of the 5 judges Bookmark placements. Figures 1 and 2 show variation in the total judge ratings but there is remarkably close similarity to the average total ratings considering that there are different judges in the two standard setting groups.

Standard Setting Procedure (Round1-2)	**Standard Region (+/-2 Std Errors)**	**Mean SME Rating**	**Median SME Rating**	**N Judges**	**Standard Error for Judges**	**Estimated Standard Error of Measurement**	**Reliability Judge Ratings**
Modified Angoff (R1)	53 to 71	62	70	8	4.57	6.58	0.74
Modified Angoff (R2)	82 to 87	85	90	8	3.91	4.77	0.81
Bookmark (R1)	70 to 92	81	80	5	10.89		
Bookmark (R2)	73 to 88	81	80	5	2.54		

Figure 1. Total Angoff Ratings by Judge Round Two (120 ITEMS)

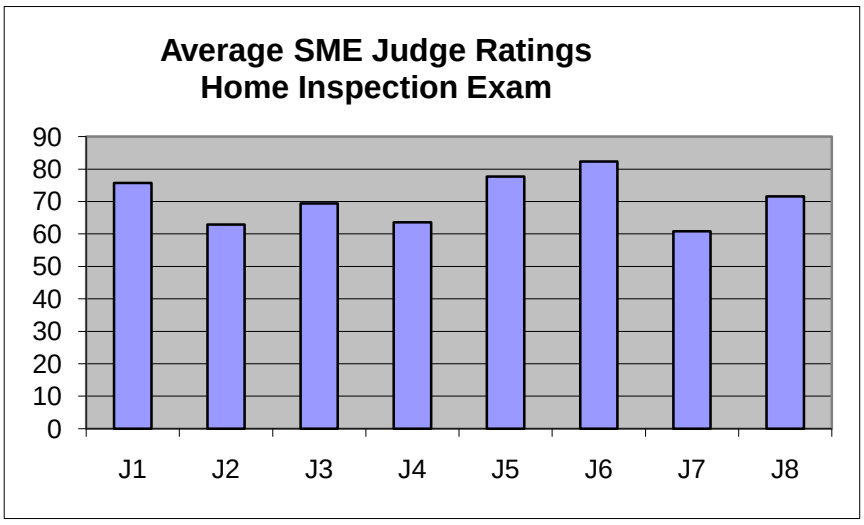
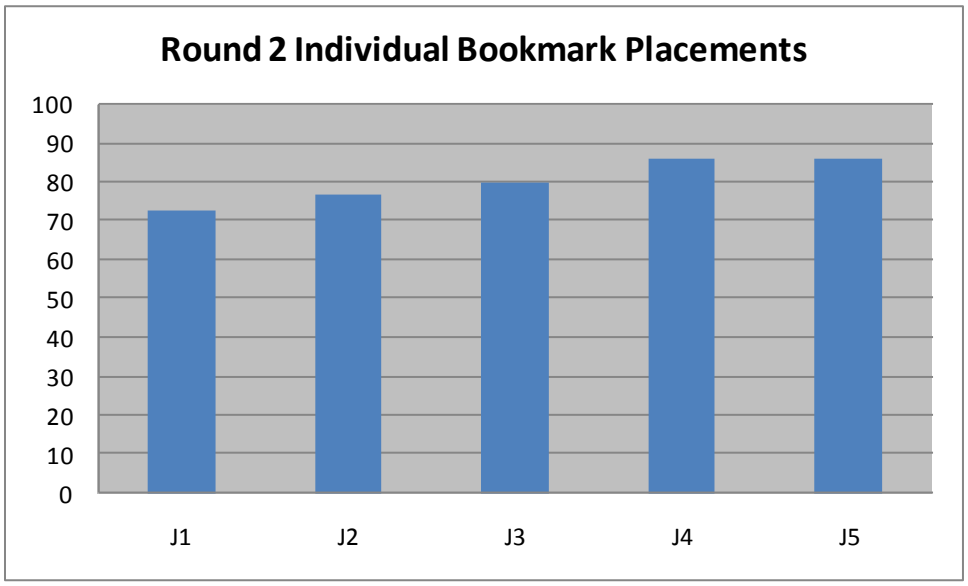


Figure 2. Individual Bookmark Placements for Home Inspection Exam – Round Two (120 ITEMS)



DISCUSSION

The results from this study provide evidence of similar cut scores being set with the Modified Angoff and Bookmark standard setting procedures. This finding is also supported by research of Buckendahl, C., Smith, R., Impara, J. & Plake, B. (2000, 2002), Davis, Buckendahl, Chin and Gerrow (2008) and Shulz (2006).

The Bookmark ratings were consistently below the results from the Modified Angoff ratings. This is similar to other research findings. Perhaps this result is due in part to the Bookmark standard judgment being computed from the item ability level estimated from the item before the judge's Bookmark placement. A recommendation was made by Reckase (2006) to compute the Bookmark cut score from an average of the Theta ability levels of the item immediately preceding and immediately following the Bookmark placement. Research reported by Shulz (2006) with the Modified Angoff and Mapmark (a variation of the Bookmark) showed that there was no systematic bias for the Mapmark (Bookmark variation) method.

This research study did not specify a specific response probability (RP=0.50 or RP=0.67) for the Bookmark approach instead examinees were asked to specify the Bookmark positioning at the point there the examinee would transition from a string of mostly correct answers to a string of mostly incorrect answers. For the Bookmark procedure the items were ordered based on Rasch model scaling; thus the RP value of 0.50 was likely similar to the personal criterion that the judges used based on the item judging procedures. Research variations with this study approach could examine variations in the implicit or explicit RP values specified for the judges (0.50, 0.60, 0.67, 0.70, and 0.75) and the effects of these variations of the RP values on the standards setting.

The investigation of multiple standard setting procedures for a given context is currently an issue showing diversity of opinion among measurement specialists. Several measurement specialists recommend use of multiple standard setting procedures to allow for comparison or triangulation of results. Other measurement specialists recommend that use of multiple standard setting procedures may just add to the complexity and cumbersomeness of standard setting, add costs to stakeholders and certification or educational organizations and that the standard setting procedure selected should be tailored to the requirements of each context.

Research should be conducted to determine the cognitive complexity of the judgmental tasks required by the judges evaluating the probability of a minimally competent candidate, a population of minimally competent candidates or a Yes/No judgment for each item with the Modified Angoff standard setting procedure. Likewise, research is needed to determine the cognitive complexity of the Bookmark task when judges are asked to judge different implicit and explicit response probabilities associated with the placement of their Bookmarks (e.g., 50%, 60%, 67%, 70%, and 75%).

One significant difference in the standard setting studies reported herein is that the Modified Angoff standard setting was conducted with an on-site face-to-face model while the Bookmark standard setting was conducted via a remote phone and web-conferencing system. There are unique strengths and weaknesses to both the on-site and remote conferencing systems. Research is needed to determine what is added and what is missing with the onsite and remote standard setting research settings. As audio and web-conferencing systems continue to add capabilities and improve quality the currently existing differences between on-site and remote site investigations will be diminished. However, there are important cues from body language and communication exchanges that cannot be captured or accommodated with remote research administrations. On the other hand, there are many advantages to conducting standard setting via a remote web-conferencing system such as reduced costs, improved accessibility to sampling judges, potential for larger representative judge samples.

Following is a summary comparison of the two standard setting studies reported herein according to the standard setting criteria recommended by Hambleton (2001). Weakness of this study are represented by the No in the columns below for the low judge panel sizes, lack of multiple panels to check generalizability, and the lack of a standard setting evaluation.

Standard Setting Criteria (Hambleton, 2001)	**Angoff for this study**	**Bookmark for this study**
Is the method for selecting judges defensible?	Yes	Yes
Are there sufficient numbers of judges both to ensure that the panel is representative of expert opinion in the field being studied, as well as to ensure that no particular judge's scores unduly influence the cut score that is set?	No	No
Will two panels be use to check the generalizability of the performance standards?	No	No

Will sufficient resources be allocated to carry out the study properly?	Yes	Yes
Will the method be field tested in preparation for use in the actual cut score setting study?	Yes	Yes
Is the cut score setting method appropriate for the particular assessment?	Yes	Yes
Will panelists be explained the purpose of the assessment and use(s) of the test score at the beginning of the process?	Yes	Yes
Will a moderator be used to help the judging panel discuss and reconcile differences?	Yes	Yes
Will the process run efficiently?	Yes	Yes
Will test data or impact data be introduced to the judgmental process so that judges can observe how their cut scores behave in practice?	Yes	Yes
Will the qualifications and other pertinent panelist data be collected?	Yes	Yes
Will the judges take the test before the standardizing procedure begins so that they have a better understanding of what examinees experienced when they [take] the test?	Yes	Yes
Will judges be trained in the method so that they have a clear understanding of their objectives and the proper process they are to follow?	Yes	Yes
Will the judging panel develop clear descriptions of the behaviors associated with each category of proficiency?	Yes	Yes
Will the approach for arriving at final performance standards be clearly described?	Yes	Yes
Will an evaluation of the process be carried out by the judges?	No	No
Will validity evidence be gathered? What form will it take?	Yes	Yes
Will the full standard setting process be documented?	Yes	Yes
Will effective steps be taken to communicate the performance standard?	Yes	Yes

The research reported herein was conducted in two separate studies with no counterbalancing, no common judges, no sampling or grouping of judges into groups, and no fully crossed Solomon Four Group research design. At the present we can only hypothesize what the effects would have been if the research study had involved a group of 16-20 judges that were divided into two panels of 8-10 judges each and each panel participated in each standard setting method in a counterbalanced and fully crossed research study. At the present time, the authors are not aware of any other standard setting study that has examined the Modified Angoff and Bookmark standard setting procedures with these research design considerations.

Generalizability studies are also needed in standard setting to determine the variance contributed by various research design facets (number of panelists, number of panels, type of content, item types, method of standard setting, complexity of the judgment task).

In Plake's (2007) Career Award Address she noted thirteen areas where further research was needed in standard setting. These research recommendations focused on design and implementation issues regarding standard setting including:

1) the effects of retakes or highest score banked,

2) use of conjunctive or compensatory scoring models,

3) effects of including stakeholders in the standard setting,

4) impact of "should" vs. "would" statements on standard setting,

5) impact of first-hand knowledge of examinee and the test context,

6) effect of taking the test in a pseudo standardized administrative conditions,

7) effect of use of subtest of items,

8) effects of training features,

9) effects of precision levels in judge ratings,

10) impact of multiple performance category settings,

11) order effects of impact information,

12) effects of 2 or 3 review rounds, and

13) effects of question phrasing on evaluation results.

She notes, "With research to support standard setters, we will be better able to make informed decisions when designing and implementing standard setting studies." (Plake, 2007, p. 21)

Hopefully the next five years of research on standard setting will yield new information on issues noted herein.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) *Educational Measurement, 2nd Edition*, Washington, DC: American Council on Education, pp. 508-597.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Brandon, P. R. (2002). Two versions of the contrasting-groups standard setting method: A review. *Measurement and Evaluation in Counseling and Development, 35*, 167-181.

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

Buckendahl, C. W., Smith, R. W., Impara, J. C., and Plake, B. S. (2000). A comparison of Angoff and Bookmark standard setting methods. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Chicago, IL: October 25-28, 2000.

Buckendahl, C., Smith, R., Impara, J. & Plake, B. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement 39*(4), 253-263.

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

Cizek, G. J. (2001a). Conjectures on the rise and fall of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Erlbaum Associates, Publishers, pp. 3-17.

Cizek, G. J. (2006). Standard setting. In Steven M. Downing and Thomas M. Haladyna (Eds.) *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, pp. 225-258.

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

Cizek, G. J., Bunch, M. B., and Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practices, 23(4),* 31-50.

Colton, D. A., and Hecht, J. T. (1981, April). *A preliminary report on a study of three techniques for setting minimum passing scores*. Symposium presentation at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.

Cross, L. H., Impara, J. C., Frary, R. B., and Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the national teacher examination. *Journal of Educational Measurement, 21(2),* 113-129.

Green, D. R., Trimble, C. S., and Lewis, D. M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practices, 22(1),* 22-32.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, pp. 80-123.

Hambleton R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Eds.) Setting performance standards: Concepts, methods and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, pp 89-116.

Hambleton, R. K. and Pitoniak, M. J. (2006). Setting performance standards. In Robert L. Brennan (Ed.). *Educational Measurement, 4th Edition*. Westport, CT: American Council on Education and Praeger Publishers. pp. 433-470.

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

Jaeger, R. M. (1989, 1993). Certification of student competence. In Robert L. Linn (Ed.) *Educational Measurement, 3rd Edition*. New York: American Council on Education and Macmillan Publishing Company; Phoenix, AZ: American Council on Education and Oryx Press. pp. 485-514.

Karantonis, A. and Sireci, S. G. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practices*, 25(1), 4-12.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement, 17(3),* 167-178.

Lewis, D. M., Mitzel, H.C., and Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), IRT-based standards-setting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers Annual Conference on Large-Scale Assessment, Phoenix, AZ.

Lin J. (2003). The bookmark standard setting procedure: Strengths and weaknesses. Paper presented at the annual conference of the Canadian Society for the Study of Education. Edmonton, CA: The Center for Research in Applied Measurement in Education. Accessed at http://www.education.ualberta.ca/educ/psych/crame/research.htm

Meara, K. P., Hambleton, R. K., and Sireci, S. G. (2001). Setting and validating standards on professional licensure and certification exams: A survey of current practices. *CLEAR Exam Review, 12(2),* 17-23.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. *Journal of Educational Measurement, 20(3),* 283-292

Mills, C. N. and Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. *Applied Measurement in Education, 1,* 261-275.

Mitzel, H. C. (2005). *Consistency for state achievement standards under NCLB*. Washington, D.C.: The Council of Chief State School Officers. P. 6.

Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In Cizek, G. J. (Ed.), *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, pp. 249-281.

Norcini, J. J., Lipner, R. S., Langdon, L. O., and Streeker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24(1),* 56-64.

Plake, B.S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education, 11(1),* 65-80.

Plake, B. S. (2005). Setting performance standards: Issues, methods. In Brian S. Everitt and David C. Howell (Eds.) *Encyclopedia of Statistics in Behavioral Science,* Volume 4, pp. 1820-1823.

Plake, B. S. (2007, April). Standards setters: Stand up and take a stand. 2006 Career Award address presented at the annual meeting of the National Council on Measurement in Education, Chicago, Il, April 2007.

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 Practices, 25(2),* 4-15.

Ricker, K. L. (2006). Setting cut scores: Critical review of Angoff and modified Angoff methods. *Alberta Journal of Educational Research, 52 (1),* 53-64.

Schultz, E. M. (2006). Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practices, 25(3),* 4-13.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.) *A guide to criterion referenced test construction.* Baltimore, MD: Johns Hopkins University Press, pp. 169-198.

Sireci, S. G. and Biskin, B. J. (1992). Measurement practices in national licensing examination programs: A survey. *CLEAR Exam Review, 3(1),* 21-25.