Item writers as standard setting panelists: Does prior experience matter?

Susan Davis-Becker

Alpine Testing Solutions

Patricia Crum

Millard Public Schools

Please address all correspondence concerning this paper to:
Susan Davis-Becker
Alpine Testing Solutions
Susan.DavisBecker@Alpinetesting.com

Abstract

A number of guidelines exist for the selection and training of standard setting panelists. Testing programs use these guidelines to convene a panel of experts who are qualified to make judgments about the expected level of performance by the population for which the exam was designed. When selecting members of this panel, program leaders may have to invite individuals who have already participated in an earlier part of the test development process (e.g., writing or reviewing items) due to the size of the potential pool of experts. The purpose of this study was to explore the impact of standard setting panelists' prior experience as item writer or reviewer on their standard setting judgments. The data were collected from eight standard setting studies conducted for a large school district. Overall, there were not any meaningful differences detected based on whether panelists had prior experience as an item writer or reviewer.

Item writers as standard setting panelists: Does prior experience matter?

Many common standard setting methods employ the use of subject matter experts (panelists) to provide judgments about the expected performance of examinees. In the Angoff (1971) methodology for example, panelists are asked to make item level judgments considering the difficulty of the item and the expectations for an examinee with a particular level of knowledge and skills. Regardless of particular variation, the task presented to the panelists can be complex (e.g., Skorupski, 2012). Therefore, selection of subject matter experts to serve as standard setting panelists must be done with careful consideration.

It is well established that the panel of subject matter experts selected for a standard setting can influence the outcome of the process. According to Hambleton and Pitoniak (2006), "Panelists' judgments are also the cornerstone on which the resulting performance standards are based" (p. 435). With this mindset, research on standard setting has led to a number of guidelines to aid in the selection of panelists. First, panelists should have knowledge of the content area and examinee population (AERA, APA, & NCME, 1999; Jaeger, 1991; Kane, 1994; Raymond & Reid, 2001) as this provides the foundation on which panelists make their recommendations. Second, panelists should have some understanding of role or performance level for which the standard is being set (Jaeger, 1991) and have an understanding of the consequences of the standards (Raymond & Reid, 2001). Third, the panelists should be capable of estimating item difficulty for the examinee population (Cizek, 1996; Hambleton, 2001; Raymond & Reid, 2001), at least in a relative sense, by utilizing their understanding of the content area and ability of the population. Finally, the full panel of panelists should represent the various stakeholder groups who will use and interpret the examination and its outcome (Hambleton, 2001; Kane, 1994; Livingston & Zieky, 1982; Raymond & Reid, 2001). Proper panel selection contributes to what Kane (1994; 2001) refers to as procedural validity evidence which is one of three types of validity evidence recommended as support the outcomes of a standard setting process.

Utilizing these guidelines, testing organizations are often tasked with finding volunteer subject matter experts who are willing to give their time to serve as standard setting panelists. Depending on the nature of the program, the population of subject matter experts who are qualified, willing, and able to serve in this role may be limited. In addition, this same population of experts is likely

also recruited for other components of the test development process (e.g., content domain analysis, item writing, item review). Therefore, in some cases, the same subject matter experts may be asked to participate in multiple test development steps. This experience can be beneficial in that 'veteran' subject matter experts can build an understanding of the test development process and major concepts (e.g., validity) which can help in their role in successive test development steps.

However, it is possible that experience in one step of the test development process could inadvertently impact a subject matter expert's efforts in subsequent steps. For example, guidance on panelist selection for alignment studies suggests that subject matter experts making such judgments should be independent of the item development process so that they do not utilize their prior knowledge of the item (intended cognitive level, content match) to make these judgments (e.g., Webb, 1999). This is an important consideration as there are often unintended characteristics of items that influence their alignment characteristics, and alignment is best evaluated by content experts who are reviewing the test content for the first time.

Thinking critically about the task faced by standard setting panelists, it should be considered that this suggestion for selection of alignment panelists may be applicable. Panelists are asked to utilize their expertise of the content area and knowledge of abilities of the intended testing population as a foundation for making judgments as to the expected level of performance. As Loomis (2012) states in the context of panelist qualifications for NAEP standard setting "these qualifications serve as the starting point for training; these credentials are intended to level the playing field for panelists" (p. 119). Therefore, for standards setting methods that involve item-level judgments (e.g., Angoff, Bookmark), the training portion of the standard setting activities includes a discussion of test and item construction and evaluation, specifically focused on some of the item-level features that can influence difficulty (e.g., content, cognitive complexity, clarity of stem, distractibility of response options) (e.g., Reid, 1991). With such training, research has shown that standard setting panelists are able to better estimate relative item difficulty as compared to absolute item difficulty.

Although each standard setting panelist is using his/her own interpretation of this training to make judgments, it is assumed that everyone is applying the same key principles to estimate item

difficulty/examinee ability relative to the item, as this is not an expected skill of standard setting panelists (e.g., Raymond & Reid, 2001). However, it is possible that panelists who participated in item writing or review may be utilizing their prior knowledge rather than making their own prediction of empirical difficulty. Raymond and Reid (2001) suggest that standard setting panelists who have served as item writers will be more aware of the particular factors that influence item difficulty due to the training they received during the item writing process. Specifically it is possible that item writers or reviewers who have prior knowledge of an item's intended level of difficulty may let this knowledge influence their judgments as to the empirical challenge for examinees and therefore, their recommended cut scores. Similarly, there can be unintended characteristics of items that influence empirical difficulty. Raymond and Reid (2001) note that the standard setting judgments of those panelists who were new to the test development process in their program were similar to those of the panelists who had served as item writers. However, this limited reference to anecdotal evidence appears to be the only consideration of test development experience in the standard setting literature.

In our experience leading standard settings, we have encountered situations where panelists have recalled the intended item difficulty or target to a particular performance level during a standard setting meeting. Although such conversations or information can be mitigated by a trained facilitator, it is unclear if such individuals are using that information to systematically inform their standard setting judgments. Therefore, the purpose of this study was to empirically investigate the impact of experience as an item writer (or reviewer) on standard setting judgments in an educational testing setting.

Methods

Assessment Program

This study includes data from eight standard setting settings conducted within a large school district. Each study was focused on a different district-wide assessment that had varying levels of stakes for students (e.g., passing a class, growth model measures impacting salary bonuses, high school graduation requirement, comparisons made between district schools), teachers, and schools. Each assessment included 55-79 multiple-choice items. Items were written to align to a specific content standard within the test blueprint and align to a particular performance level to

ensure that the full test covered the range of expected performance. Four of the studies were focused on setting cut scores for four performance levels (three cut scores: level 2, level 3, level 4) and four studies were focused on setting cut scores for three performance levels (two cut scores: level 2, level 3).

During the item writing process, items were written with a targeted content and performance level (i.e., conceptual difficulty) match. During the standard setting process, the subject matter experts were asked to judge the expected performance of students who were at the transition point from one performance level to the next (e.g., *Just* Proficient, *Just* Advanced) in two rounds of standard setting. Therefore, knowledge of the intended performance level match could be directly applied to such standard setting judgments. Across the execution of these eight standard setting studies, facilitators observed that occasionally a subject matter expert would reference knowledge of the intended item content or conceptual difficulty match. This anecdotal evidence suggests that, at least some, subject matter experts who participated in both item writing and standard setting activities were able to recall information during the standard setting from the earlier process.

Standard Setting Panelists

For each standard setting study, the district assessment office recruited subject matter experts with the assistance of district content and grade level curriculum facilitators. These facilitators were familiar with the district teachers as they trained all of them in instructional strategies and worked with groups of them to develop and align district standards and curriculum. The qualifications for item writing and standard setting participants were essentially the same. The teachers selected were considered to be the most knowledgeable about both content and instructional methodology in the district. In addition, the students they teach represented the district demographics, including the range of geographic locations within the district as well as student populations (general education, special education, and English language learners). For Middle level and high school the selection also included proportional representation of student course enrollment (e.g., honors, regular and basic courses in mathematics). Because of the limited population of subject matter experts within these districts, the same individuals were often recruited for multiple test development steps.

For each of the standard setting studies, the district recruited between 15 and 24 panelists (average panel size was 20). Across the studies, 20 to 75 percent of these panelists had prior experience as an item writer or item reviewer for that particular assessment. For each study, the panelists were identified as part of a focal group (item writers, item reviewers, either) or the non-item/reviewer writer group. Group level differences in recommended cut scores were examined for Round 1 and Round 2 judgments within and across studies.

Standard Setting Process

The data included in this study were gathered over eight standard setting studies that used the Yes/No modification of the Angoff (1971) method (Impara & Plake, 1997). Each standard setting was conducted following the same process. The panelists participation in a training and orientation at the beginning of the day where they reviewed the purpose of the assessment, the assessment specifications (i.e., content framework, weighting), and the policy descriptors for each performance level. At this point in the process, the panelists were asked to use the policy descriptors for each performance level and review the assessment specifications to define the expected knowledge and skills of the "target students" - those who were at the transition point from one performance level to the next (e.g., Barely Level 2, Barely Level 3, Barely Level 4). These expectations were recorded and copies were provided to the panelists for the remainder of the standard setting activity.

To begin the Angoff standard setting activity, the panelists were provided additional training on the method and what types of ratings they were asked to provide including a review of the factors that can contribute to item difficulty. This part of the training encouraged panelists to think about the conceptual challenge of the item (e.g., content, cognitive processing required) as well as specific item-level features that can make an item easier or more difficult (e.g., question phrasing, clarity of the distractors, clues present in the item). At this point, the panelists were given an opportunity to practice making Angoff ratings with a sample of ten items. Specifically, they were asked to estimate which of the target students would likely answer each item correctly. These items were drawn from the same pool as the operational form and were of varying levels of difficulty. After making their initial ratings, the panelists were guided in a discussion where they could share their rational for their judgments with the panel.

The operational component of the Angoff studies was conducted in two rounds. During the first round of ratings, panelists were asked to make their ratings relying on their knowledge of student ability, the descriptions of the target students created earlier in the standard setting meeting, and their assessment of item difficulty using the training provided. After the first round of ratings, panelists were provided feedback including (1) their individual recommended cut scores, (2) the panel's recommended cut scores including measures of variability, and (3) the impact of the panel's recommended cut scores in the form of percent of students classified within each performance level, and (4) the item p-values as a measure of empirical difficulty of each item. After a brief discussion of the feedback information, panelists were asked to make their second and final ratings for each item. This second round of ratings was used to formulate the final recommended cut scores.

Results

Analysis of Recommended Cut Scores across Studies

Across the eight standard setting studies, there were a total of 157 panelists and 20 recommended cut scores. As a first step in this analysis, all cut scores (e.g., individual recommended cut scores, group-level cut scores) were converted to percentage values (of the total test score). This was done to allow for comparison of results across studies as each exam was of a different length. Panelists were identified as having served as an item writer or reviewer (experience) or as new to the test development process (no experience). Information about each standard setting study is provided in Table 1 including number of multiple choice items on the assessment, the number of cut scores that were a part of the study, and the percentage of the panel who had test development experience with that particular assessment.

Table 1. Standard setting panelist counts, item counts, cut score counts, and portion of panelists with prior test development experience

Study	Panelists	Items	Cut Scores	% of Panel with Experience[1]
1	23	79	2	30%
2	16	64	2	75%
3	21	59	2	38%
4	18	65	2	72%
5	24	56	3	46%

6	20	62	3	30%
7	20	71	3	20%
8	12	55	3	27%

[1]Experience here is defined as experience as an item writer or reviewer for the specific test that is the focus of the standard setting study

It was expected that any differences between panelists who had prior knowledge of the test content and those who would not likely appear during the first round of ratings, before data was provided as to the empirical difficulty of each item. However, our inspection of the results included both rounds of judgments. The average cut score, by study, by cut score level, by round, and by group (experience, no experience) is presented in Table 2. Comparing the average recommended cut scores between groups across levels and studies suggested that the difference between groups ranged from 0-10% in round 1 and 0-11% in round 2. Averaging across all studies and cut scores, the differences within round and between groups were less than half a percent suggesting that there were no systematic differences between the cut scores recommended by each group.

Table 2. Recommended cut scores by study, round, level, and group

		Round 1		Round 2	
Study	**Cut Score[1]**	**No Experience Cut Score (%)**	**Experience Cut Score (%)**	**No Experience Cut Score (%)**	**Experience Cut Score (%)**
1	L 2	52%	47%	57%	53%
1	L 3	90%	91%	89%	94%
2	L 2	32%	41%	45%	49%
2	L 3	84%	88%	77%	84%
3	L 2	47%	45%	36%	47%
3	L 3	82%	81%	78%	79%
4	L 2	48%	54%	58%	53%
4	L 3	82%	90%	85%	81%
5	L 2	30%	30%	26%	26%
5	L 3	65%	75%	65%	64%
5	L 4	95%	98%	93%	91%
6	L 2	38%	35%	34%	30%
6	L 3	86%	76%	67%	63%
6	L 4	100%	99%	95%	94%
7	L 2	43%	42%	43%	40%
7	L 3	77%	74%	76%	73%
7	L 4	93%	93%	93%	94%
8	L 2	39%	31%	42%	48%

Study	Cut Score[1]	Round 1 No Experience Cut Score (%)	Round 1 Experience Cut Score (%)	Round 2 No Experience Cut Score (%)	Round 2 Experience Cut Score (%)
8	L 3	73%	70%	79%	76%
8	L 4	94%	95%	97%	92%

[1]L2 = Level 2 cut score, L3 = Level 3 cut score, L4 = Level 4 cut score

To compare the variability in the cut scores, the standard error was estimated for each study, for each cut score, for each round of standard setting judgments, and for each group. Kane (1994; 2001) has identified these values as one source of internal validity evidence as they can provide a measure of the agreement among members of the panel. These results are presented in Table 3. Although there are some differences within specific studies and cut scores, these are sporadically found across studies, cut scores, and rounds. In addition, the average standard error presented for each group across all cut scores within a given round suggests that any systematic differences (i.e., across cut scores and studies) were less than 1% of the total test score.

Table 3. Standard error estimates by study, round, level, and group

Study	Cut Score	Round 1 No Experience	Round 1 Experience	Round 2 No Experience	Round 2 Experience
1	L 2	2%	3%	2%	2%
1	L 3	1%	2%	1%	1%
2	L 2	6%	2%	3%	2%
2	L 3	7%	2%	11%	4%
3	L 2	2%	4%	3%	4%
3	L 3	2%	2%	4%	3%
4	L 2	5%	3%	10%	3%
4	L 3	5%	2%	7%	4%
5	L 2	3%	2%	3%	3%
5	L 3	4%	3%	3%	5%
5	L 4	2%	1%	1%	3%
6	L 2	2%	2%	3%	3%
6	L 3	3%	2%	2%	5%
6	L 4	<1%	<1%	1%	3%
7	L 2	2%	3%	2%	3%
7	L 3	2%	7%	2%	6%
7	L 4	2%	3%	1%	4%
8	L 2	3%	6%	2%	3%
8	L 3	3%	6%	3%	5%
8	L 4	2%	4%	2%	2%
AVERAGE		3.1%	3.1%	3.3%	3.4%

The goal of convening an appropriate standard setting panel is that if the study was replicated with another qualified panel, the results would not be meaningfully different. To determine if any of the differences in recommended cut scores between groups (Table 2) were meaningful, the recommended cut score range from each group was estimated by using the mean $\pm$ 1 group-level standard error. For the purposes of this comparison, the cut scores for the two groups were identified as being notably different if the recommended cut score ranges did not overlap. The results of this analysis are presented in Table 4. Most ranges were overlapping with a few exceptions (four in round 1 and two in round 2, identified in bold text in the Table). These exceptions seemed minor and did not suggest any pattern as they were varied across studies and across the cut scores (2 or 3). Out of the six non-overlapping cut score ranges, five were instances where the Experience group recommended a higher cut score than the No Experience group. In addition, the finding that there were four recommended cut score ranges that did not overlap in Round 1 and two in Round 2 might suggest that any differences that did exist were reduced after the feedback was introduced.

Table 4. Comparison of recommended cut score ranges ($\pm$ 1 SE) by study, cut score, round, and group

		Round 1		Round 2	
Study	**Cut Score**	**No Experience**	**Experience**	**No Experience**	**Experience**
1	L 2	50% - 53%	44% - 51%	55% - 58%	51% - 55%
1	L 3	89% - 91%	89% - 92%	**87% - 90%**	**93% - 95%**
2	L 2	**27% - 38%**	**39% - 43%**	42% - 47%	47% - 51%
2	L 3	77% - 91%	86% - 90%	66% - 87%	80% - 88%
3	L 2	45% - 50%	41% - 49%	43% - 49%	43% - 51%
3	L 3	80% - 84%	79% - 83%	74% - 82%	76% - 82%
4	L 2	43% - 53%	50% - 57%	48% -67%	50% - 57%
4	L 3	**78% - 87%**	**88% - 92%**	78% -92%	77% - 85%
5	L 2	27% - 32%	28% - 32%	23% -28%	23% - 29%
5	L 3	**62% - 69%**	**71% - 78%**	62% -68%	59% - 70%
5	L 4	94% - 97%	97% - 99%	92% -94%	88% - 94%
6	L 2	37% - 40%	33% - 38%	31% -37%	27% - 33%
6	L 3	**83% - 89%**	**74% - 77%**	65% -69%	58% - 68%
6	L 4	100% - 100%	99% - 100%	93% -96%	91% - 97%
7	L 2	42% - 45%	39% - 45%	40% -45%	37% - 43%
7	L 3	75% - 79%	67% - 82%	74% -78%	67% - 79%
7	L 4	92% - 95%	89% - 96%	92% - 94%	91% - 98%

		Round 1		Round 2	
Study	**Cut Score**	**No Experience**	**Experience**	**No Experience**	**Experience**
8	L 2	36% - 42%	26% - 37%	**39% - 44%**	**45% - 52%**
8	L 3	70% - 77%	64% - 75%	76% - 82%	71% - 81%
8	L 4	93% - 96%	91% - 99%	95% - 99%	90% - 94%

Analysis of item level judgments from one study

To further explore the possibility of differences, item-level judgments from one particular study (study #8) were analyzed. As was shown in Table 4, the recommended cut score ranges were overlapping for all but one comparison (Level 2, Round 2).

Although there were no systematic differences in the recommended cut scores by group, we next asked whether the Experience group would be better at estimating relative item-level difficulty than the No Experience group. Although there is no "true" cut score or expected impact distribution, it would be expected that standard setting panelists would make their ratings as to correspond with the empirical item difficulty (e.g., Reid, 1991). Specifically, the Angoff judgments made by the panelists were done in a way that they classified items into one of four categories:

- Likely to be answered correctly by a student who is Barely Level 2
- Likely to be answered correctly by a student who is Barely Level 3
- Likely to be answered correctly by a student who is Barely Level 4
- None (Not likely to be answered correctly by any of the target students)

To investigate this, we computed the average empirical item difficulty for items classified within each of the four categories listed above. There was an expected relationship between category and average item difficulty in that the higher levels would have more difficult items (e.g., the p-values in the "Barely Level 4" category would be lower than those in the "Barely Level 3" category). These results are shown in Table 5. It was expected that any effect of experience would be most apparent in the Round 1 rates. However, the results of the first round of ratings only show differences in the two highest categories (Barely Level 4, None) and these results show a lack of differentiation among the average difficulty of items within each category as both groups classified items within the "none" category that were, on average, easier than the average

item within the Barely Level 4 category. This finding is likely due to the smaller number of items on the assessment identified as targeted towards this end of the ability range. The results for Round 2 suggest that both groups of Panelists (No Experience, Experience) utilized the item difficulty information provided as feedback to adjust their ratings so that the most difficult items were classified in the "none" category.

Table 5. Average p-value for items classified as likely to be answered correctly by students at each performance level

	Round 1		Round 2	
	No Experience	**Experience**	**No Experience**	**Experience**
Barely Level 2	0.93	0.93	0.95	0.95
Barely Level 3	0.87	0.87	0.86	0.85
Barely Level 4	0.81	0.84	0.81	0.80
None	0.84	0.87	0.70	0.69

Conclusions

With appropriate training, Angoff-type ratings provided by qualified panelists should reflect the relative difficulty of test items for the target examinees (Reid, 1991). However, panelists could potentially let other knowledge influence their ratings with or without knowing. This confound could limit the validity of the results if there were to be evidence that panelists were allowing this background knowledge to influence their estimates of examinee performance. The purpose of this study was to empirically investigate the influence of panelists' prior item writing/reviewing experience on standard setting judgments within one district program.

It was expected that any differences due to experience would be most evident in the Round 1 ratings. The initial investigation of the Round 1 recommended cut scores revealed some differences between groups but no clear pattern across studies and cut scores. Similarly, the standard errors estimated for each performance group indicated some differences but nothing systematic. To determine if either of these values (recommended cut scores, standard errors) amounted to meaningful differences between groups, the recommended cut score ranges were estimated for each group and within each study for each cut score. Within Round 1, out of twenty recommended cut scores, four had recommended ranges by group that were not overlapping. In three of these four cases, the cut scores were from the standard setting studies

with the largest percentage of the panel being experienced. However, if the percent of the panel with experience was an influencing factor, we would have expected to see such differences in the other cut scores within these studies.

We also reviewed the item-level judgments for one study in particular to determine if any differences could be found when looking at this finer level. The item-level judgments were evaluated using the empirical item difficulty information (p-values). Even at this level, there did not appear to be any meaningful differences in the ratings provided by the two groups. In fact, both groups appeared to use item difficulty information provided as feedback to better align their item-level ratings in their second round of judgments. This finding parallels what was concluded from the panelist evaluation results across these studies – the panelists reported that they found the item difficulty information to be the most useful piece of feedback provided (as compared to their recommended cut scores, the panel's recommended cut score and associated impact, and the panel discussion between rounds).

If the experienced panelists had some premonition of what the *intended* difficulty of an item might be, the empirical item difficulty may have been higher or lower than expected and they may have misjudged in both directions in their initial judgments. However, regardless of experience, all panelists seemed to make their final ratings (which were used to estimate the recommended cut score) by utilizing the empirical item difficulty information.

There were several limitations to this research that are worth noting. First, the standard setting studies included were all from one particular assessment program and the results may not generalize to other programs. Second, given that each study included three or four performance levels, panelists' ability to recall information about the items (i.e., intended level of difficulty) may have been confounded by the number of possible classifications. Panelists may have had the ability to recall additional information had they only been focused on two performance levels. Third, although panelists were able to make their own independent judgments, whenever the subject of "intended difficulty" came up during a standard setting conversation, the facilitator reminded the panelists that they were to make their judgments as to the actual difficulty of the items (rather than the intended difficulty). Therefore, they may have intentionally focused on the item characteristics (as described in the standard setting training) rather than prior knowledge.

In summary, the findings from this empirical investigation suggest that there are no meaningful differences in the standard setting judgments between item writers/editors and non-item writers/editors. In turn, this evidence can be used as justification for allowing veteran subject matter experts to participate as standard setting panelists. Given the importance of standard setting to the test development process, future research will likely continue to explore panelist characteristics that contribute to a defensible standard setting process and results.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement,* (2nd ed., pp. 508-600), Washington, DC: American Council on Education.

Cizek, G. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice, 15(1),* 13-21.

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.

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.

Impara, J. C., & Plake, B. S. (1997). An alternative approach to standard setting. *Journal of Educational Measurement*, 34(4), 355-368.

Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10(2),* 3-10.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64(3)*, 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, N.J.: Educational Testing Service.

Loomis, S. (2012). Selecting and training standard setting participants: State of the art policies and procedures. In G. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (2nd Edition, pp. 107-134). New York, NY: Routledge.

Raymond, M. & Reid, J. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. Cizek (2001). Setting performance standards: Concepts, methods, and perspective. Mahwah, NJ: Lawrence Erlbaum.

Reid, J.B. (1991). Training judges to generate standard setting data. *Educational Measurement: Issues and Practice, 10(2),* 11-14.

Skorupski, W.P. (2012). Understanding the cognitive process of standard setting panelists. In G. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (2nd Edition, pp. 135-148). New York, NY: Routledge.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 18). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.