Measuring the Power of Learning.™

Research Memorandum
ETS RM–19-12

Standard Setting Panelist Cognition: A Framework and Implications for Practice

Irvin R. Katz

October 2019

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Standard Setting Panelist Cognition: A Framework and Implications for Practice

Irvin R. Katz
Educational Testing Service, Princeton, New Jersey

October 2019

Corresponding author: Katz, I. R. E-mail: ikatz@ets.org

Suggested citation: Katz, I. R. (2019). *Standard setting panelist cognition: A framework and implications for practice* (Research Memorandum No. RM-19-12). Princeton, NJ: Educational Testing Service.

Action Editor: James Carlson

Reviewers: Priya Kannan and Richard Tannenbaum

Abstract

Standard setting plays a significant role in the validity argument for assessments. Standard setting practices at Educational Testing Service (ETS) have evolved over time, motivated by practice and empirical research from the standard setting literature. However, few efforts in the literature have been directed toward understanding the drivers of standard setting decisions. By reviewing empirical research that might inform how experts make standard setting judgments—including cognitive psychology research rarely considered in the development of standard setting practice—this paper seeks both to inform current ETS practices and suggest research studies.

Key words: Angoff method, cognition and assessment, cognitive biases, cut scores, probabilistic reasoning

In 2005, a special issue of *Applied Measurement in Education* contained several studies designed to address the issue of standard setting panelist cognition. These studies were motivated by a recognition that the standard setting literature contains numerous recommendations and practical rules of thumb, but none of the recommendations are driven by detailed knowledge of what panelists think about to make their standard setting judgments. Unfortunately, these studies seemed to have little impact on the field, with few studies published on this topic since 2005 (Hein & Skaggs, 2009, 2010; Papageorgiou, 2010).

These studies seek to understand panelist cognition through qualitative methods: researchers "ask" panelists (whether individually or part of a group discussion) about their judgments or related aspects of the standard setting meeting. McGinty (2005) took notes during panelists' discussions and identified frequently mentioned difficulties. Panelists seemed particularly confused as to the policy-level decisions associated with standard setting and how they were to apply their expertise. However, McGinty's work was not focused on panelist cognition directly but on a tangential issue: the misunderstandings that can arise when panelists do not receive clear training on the standard setting methodology, such as which portions of the meeting relate to policy (i.e., what "should" be the performance of successful test takers) and which relate to domain judgments (i.e., how "would" a just-qualified candidate or borderline examinee perform on this test).

Skorupski and Hambleton (2005) asked panelists to fill out questionnaires at specific points in their standard setting study. The researchers reported panelists' evolving understanding of the performance level descriptions and the standard setting process as well as changes in panelists' confidence in their own judgment. Panelists answered other questions about the influence of study materials (e.g., description of skills required to "pass," also known as performance level descriptions), other panelists' judgments, and item difficulty data. After a second round of judgments, the panelists were asked about their strategy for making the standard setting decisions.

Ferdous & Plake (2005) conducted focus group interviews within their standard setting study to hear about the materials that influenced panelists' decisions. Using analyses of cut score judgments, researchers divided the panelists into groups: those who set relatively high (stringent) cuts, medium cuts, or low (less stringent) cuts. The researchers compared the focus group discussions across the three groups. For example, while all panelists reported paying attention to

the performance level descriptions, the people who assigned the lowest cut scores reported considering the testing requirements of the No Child Left Behind Act while those who assigned higher cuts did not consider these requirements.

Both Hein and Skaggs (2009, 2010) and Papageorgiou (2010) used a focus-group methodology similar to that of Ferdous and Plake (2005). The researchers documented difficulties in several steps of the standard setting process, including the wording of performance descriptors and how to conceptualize a group of barely qualified students.

One reason for lack of progress in this area might be that none of the prior research gets to the heart of the matter: What cognitive processes are involved in making standard setting judgments? Qualitative studies might provide clues as to what panelists find confusing, what information they attend to, and whether they appropriately interpret key study materials (e.g., the definition of the just qualified candidate [JQC] or borderline examinee [BE], the characteristics that lead to easier or more difficult items, the specifics of the standard setting methodology). However, asking panelists what they are thinking—whether via interviews, verbal protocols, focus groups, or surveys—provides no direct information into the cognitive mechanisms that sit between the inputs (e.g., JQC or BE definition, item, standard setting prompt) and the output (the panelist's rating). Although these methods might suggest finer grain mental steps that would otherwise occur silently, verbalization of such mental steps is only possible if the person is consciously aware of them (Ericsson & Simon, 1993). Needed instead is research that builds theories or models of the unobservable (and potentially unconscious) cognitive mechanisms and compares the performance of those models with human performance.

In this paper, I outline a theory of panelist cognition based on the cognitive psychology research literature on judgment and decision making and cognitive modeling. This theory is not yet fully fleshed out, but rather a direct application of a cognitive model from the judgment and decision-making literature on how people make probability estimates (see Moon, Finn, LaMar, & Katz, 2018, for an introduction to the use of computational cognitive models for assessment).

The paper is organized as follows. First, I identify the portion of a standard setting study that is the target of the paper: panelists' individual judgments and where they occur in the flow of a standard setting meeting. Next, I argue that these judgments may be interpreted as instances of probabilistic reasoning and introduce cognitive models that have been used to explain probabilistic reasoning in other contexts. Finally, I apply this cognitive framework to the

situation of standard setting and draw some implications for research and practice that arise from the model of standard setting panelist cognition.

Common Elements of Standard Setting

Standard setting is the "proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more conceivable states or degrees of performance" (Cizek, 1993, p. 100). Although a process of informed professional judgment, standard setting is now recognized to be an integral part of the validity evidence supporting test score use (Bejar, Braun, & Tannenbaum, 2007; Kane, 2006; Papageorgiou & Tannenbaum, 2016).

Panel-based standard setting approaches contain several common elements—activities that the panelists complete regardless of test content or specific standard setting methodology (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Tannenbaum & Katz, 2013). Panelists learn about the key concepts underlying standard setting. They become familiar with the test for which a cut score is to be set, typically by taking the test, self-scoring, and discussing their experiences with other panelists. Panelists typically discuss the performance level description(s) (PLD), which is the description of test-taker knowledge and skills required to achieve a particular level. Depending on the test, these descriptions might be the knowledge and skills of a proficient math student, a beginning French teacher, or a professional engineer. During the meeting, panelists work from the PLD to outline the expectations for a test taker just at the cut point, alternatively referred to as the minimally competent examinee, JQC, or BE. Panelists receive training and practice in the standard setting method and then make a first round of independent judgments. Subsequent group discussion and feedback, which may include item difficulty or classification percentages, sometimes lead panelists to change their judgments, which is done during subsequent rounds of individual judgments. The final round of judgments concludes with a decision on a recommended cut score(s), which might be determined through consensus or by averaging panelist responses.

Commonly, at two points in the meeting—after training and after making the final recommendation for the cut score(s)—panelists complete surveys that document their beliefs about the validity of the standard setting process. For example, panelists report whether they believe the instruction and training were adequate, what aspects of the meeting influenced their judgments, and whether they agree with the recommended cut score(s).

Standard Setting "Kernel:" Individual Judgments

The individual judgments have been called the "kernel" of the standard setting process (Brandon, 2004) because they are the data used to calculate cut scores. In contrast, the surrounding activities (taking the test, discussing a PLD and defining the BE, training, group discussions, and feedback) have the primary goal of supporting the kernel. That is, these activities provide evidence for the validity of the judgments: that judgments reflect panelists' content expertise and experience by minimizing various irrelevant (to the standard setting task) factors. Irrelevant factors might include a systematic misunderstanding of the test's purpose by panelists or panelists attributing more knowledge to the BE than is warranted by the PLD.

Different standard setting methods, or different implementations of the same method, require different types of judgments. However, most standard setting methods that focus on the test items (rather than on exemplar responses, such as Body of Work or Contrasting Groups; Cizek & Bunch, 2007; Zieky, Perie, & Livingston, 2008) involve a type of probability estimation. For example, in the Angoff method, a typical judgment is the likelihood (probability) that a BE—someone with just enough knowledge and skills to achieve the cut score—would answer each test item correctly, with separate probability judgments made for each item. In the Nedelsky method, panelists judge the multiple-choice response options that a BE would be able to eliminate from consideration, effectively answering the question, "Would a BE be able to eliminate this option NN% of the time?" where NN is some threshold of performance. A study using the Bookmark method might appear different on the surface: panelists review a set of items ordered by difficulty and judge the item that distinguishes between what a BE can and cannot answer correctly. However, the individual standard setting judgments are similarly a probability estimate: Would a BE be able to answer correctly a particular multiple-choice item 67% of the time?

Based on the above observations, it seems reasonable to conclude that many types of panelist cognition focus on the estimation of the likelihood of an event (e.g., a person answering an item correctly). In the field of cognitive psychology (the judgment and decision making literature in particular), research has focused on how people make probability estimates both under laboratory conditions (reviewed below) and in a variety of real-world settings. Real-world probability judgments have included predicting success of a new product (Astebro & Koehler, 2007), evaluating military air threats (Bryant, 2007), and predicting various types of fiscal events

(McCaffery & Baron, 2006). Much of this research focuses on developing cognitive models: explanatory mechanisms that describe not only the observable aspects of a situation (e.g., the information presented and the person's predictions), but also the unobservable cognitive processes that mediate the predictions. As will be discussed below, such models provide a useful framework for describing new predictive situations, such as standard setting panelists' judgments.

Cognitive Models of Probabilistic Reasoning

Several researchers have developed cognitive models of how people generate estimates—whether of numerical values (What will be the price of company X's stock next year?), preference (Is product A or B the better choice?), or probability (What's the likelihood that this patient has the flu?). Interestingly, these models describe the situation to be modeled similarly, in terms of its attributes and levels for those attributes. This common framework extends across models, suggesting a possible approach to viewing the situation of interest in this paper, that of estimating the likelihood that a BE can answer a test item correctly.

Nilsson, Olsson, & Juslin (2005) developed five mathematical models of probably judgment and conducted experiments to investigate the match between each model's prediction and human behavior. The models differed both in the cognitive processes involved and in the assumptions they made about the structure of memory. The models were developed based on theories of subjective probability judgments from the judgment and decision making and broader cognitive psychology literatures. Some models were based on alternative interpretations of the "representativeness heuristic"—the cognitive bias whereby people believe that the closer an event is to the general class of such events (the more representative), then that event is more probable. For example, if a description of a person fits the stereotype of an engineer (likes math, is a loner), then people are likely to classify that description as being of an engineer regardless of the frequency of engineers in the sample. In other words, people ignore the base rate of events in favor of the representativeness of an event or object. Three models derive from this heuristic:

- Prototype similarity. An event is more likely to be in Category A to the extent that it is similar to a prototype event. In the model, similarity is operationalized as the closeness of the features of the event (the probe) to the prototype. This model assumes that people have a prototype event in their head, with its own features and values.

- Relative likelihood (holistic): An event is likely to the extent that its features appear in Category A (vs. Category B).
- Relative likelihood (features): As the above model, except that the relative likelihood of each feature is calculated separately.

Two other models were also examined:

- Exemplar-based, one cue: The probability that an event is in Category A is more likely to the extent that the event has a critical feature that appears in other events within Category A. The assumption is that all events are stored in memory, and so the comparison is made to all these events rather than to just an exemplar. Also important is that the comparison is made based only on the most critical feature (a single cue) that tends to distinguish Category A events from other events.
- Exemplar-based, all cues: As above, except all cues/features associated with an event are considered, rather than just the most distinguishing cue. In other words, people consider all of the information they have seen about the relationship between attributes–values and their probability of occurrence.

This last model (exemplar-based, all cues) provided the best match to human data in several experiments.

von Helversen and Rieskamp (2008) used a similar framework in their investigation of estimation skill. In the researchers' framework, estimation is based on the presence or absence of cues. For example, a cell phone might or might not have a digital camera, Internet access, or an adequate display size. Note that the values of the cues (e.g., the particular megapixel rating of the digital camera, the specific size of the display) is not considered in this framework. In a typical experiment, participants learn about the cost of different objects (e.g., cell phones) each having different patterns of absent vs. present cues (or features). After learning these costs, participants then estimate the costs of novel cell phones that have different patterns of features.

von Helversen and Rieskamp (2008) developed several alternative cognitive models of the above estimation task. Generally, their results were most consistent with a model wherein the number of present cues is counted and the median cost of other items with a similar "cue count" is reported as the estimate. These results demonstrate that for some domains, people take "shortcuts" in their estimates, considering only how many features are present.

Other researchers have investigated the qualities of the cues that people tend to pay attention to when making probability judgments of estimates and the effect of that information on estimation strategies. For example, people pay attention to more salient (vs. implied) information (McCaffery & Baron, 2006), appear to be influenced more by information presented as frequencies than as proportions (Friedrich, Lucas, & Hodell, 2005), can become distracted by irrelevant information (Hall, Ariss, & Todorov, 2007), and more heavily weight cues that agree with each other (Karelaia, 2006). In addition, people sometimes adapt their decision strategy depending on how information is formatted, preferring strategies that rely on the more easily extracted information from a display (Zhang, Hsee, & Xiao, 2006).

To the extent that people ignore certain cues or focus overly much on others, it might diminish the quality of the estimates based on the cues. Even if an attribute can be objectively shown to influence some future probability, if people do not pay attention to that attribute, then their estimates might not reflect their expertise as well. Research of this type is important because it shows the biases that people bring with them to a probability estimation task and, by implication, potentially to the judgments during standard setting. To the extent that we can make panelists aware of these biases, or present the attributes of items or characteristics of the borderline examinee in a way that minimizes these biases, the quality panelists' judgments during a standard setting meeting should improve.

Application of the Cognitive Framework to Standard Setting

How does this framework apply to the standard setting situation? The two external pieces of information that should guide each probability judgment are the definition of the BE and the test item under consideration. In line with models of probability judgment, we assume that both types of information may be viewed as sets of attribute–value pairs.

The pertinent information from the BE is the list of knowledge and skills possessed (the attributes of a BE) and the level of each knowledge or skill (the values). The level of the knowledge or skill, or the context of its use, is particularly important as it defines the limits of ability for the JQC. For example, the PLD for NAEP fourth-grade mathematics Basic level reads: "Fourth graders performing at the Basic level should be able to estimate and use ***basic*** facts to perform ***simple*** computations with ***whole numbers***; show ***some*** understanding of fractions and decimals; and solve ***some simple*** real-world problems. . . . [emphasis added]" (Perie, 2008, p. 19). (Note that this is a general description of a performance level, a PLD, rather than the

narrower definition of a BE that might be developed during a standard setting meeting.) This definition includes both a description of the skill (the attributes) as well as the level of the skill possessed by the student at the basic level (emphasized words). The level of the knowledge and skills provides a basis for comparison to the knowledge and skills required by a test item.

A test item may be similarly viewed as a set of attribute–value pairs, except in this case, the description is of the knowledge and skills needed to solve the item successfully. For example, a fourth-grade mathematics item might require ***simple*** computations using ***non-whole*** numbers. A French item might require the candidate to know the meaning of a ***common*** idiom as well as ***basic*** vocabulary. In each case, the bold-italicized words represent the level of the knowledge or skill demanded by the items.

To make a standard setting judgment—the probability estimate that a BE can correctly answer a given item—the panelist compares his or her representation of the BE and the item's cognitive demands. The details of this comparison, and the extent to which it incorporates deliberate reasoning versus implicit reasoning through memory retrieval, are beyond the scope of this paper and requires further research. However, at least two potential strategies appear likely. One strategy is based on anecdotal evidence from standard setting studies (observed during group discussions of panelists' rationales for their ratings); the other strategy is based on the probability estimation modeling research reviewed above.

First, a panelist might focus on the BE definition—the specific words that describe the knowledge and skills possessed by the borderline examinee—and make judgments based on these abstract descriptions of skills. The probability of successful solution is the extent to which the definition matches the demands of the item. For example, a panelist might look at the skill level written (or implied) in a BE and see that the BE should be able to solve routine computations with whole numbers. If the task involves routine computations with decimal numbers, the panelist would estimate the degree of difficulty added by the decimal numbers (an item involving routine computation on whole numbers should be solved every time by such a BE). Critical to this strategy is that the judgments focus on the abstract descriptions of skills and levels, rather than concrete exemplar memories of student performance.

Second, in line with the cognitive models, a panelist might base his or her probability estimation on the memory of students solving problems that are like those in the test. As in the exemplar model of Nilsson et al. (2005), the probability estimate would be the ratio of the

number of successful attempts at similar items by students who are like the BE to the number of all such attempts. In this case, "similarity" might reflect the degree to which the set of attribute–value pairs for the BE overlap with those of students in the panelist's memory. In the same way, the problems attempted by students that the panelist has seen may be represented in the panelist's mind in terms of attribute–value pairs, and so "similar" items would be those whose characteristics (attribute–value pairs) overlap with those of the item currently being considered.

Interestingly, the second strategy—the one consistent with the research literature on probability estimation—requires memories and skills that appear consistent with those possessed by many teachers. Experienced teachers presumably have seen many students attempting to solve many types of problems, and so they should have a rich store of memories on which to base any probability estimation. In addition, because of their expertise, teachers should be able to recognize the demands of given test items, at least generally. Hambleton, Sireci, Swaminathan, Xing, and Rizavi (2003), in their review of the literature on teachers or item writers making estimates of item difficulty, noted that these groups are fairly proficient at ordering items by difficulty, even if they cannot estimate the exact difficulty of an item administered to a general population. However, for purposes of estimating probability in this model, being able to recognize what makes an item more or less difficult (the attribute–value pairs representing the item's cognitive demands) might be sufficient for producing a reasonable estimate of the likelihood that a BE could successfully solve the item.

This framework for understanding the cognitive processes of standard setting panelists should be explored and enhanced through further research. Research might address the extent to which the identification of cognitive level of skills, represented in the BE or discovered through inspection of items, influences standard setting ratings. Another line of research might investigate the existence of alternative strategies for making standard setting ratings. For example, evidence might be sought for the strategies outlined above (consideration of abstract skills vs. estimation based on exemplars of performance). Research might also investigate the strategies that lead to more consistent performance as well as the training and facilitation techniques that would guide standard setting panelists to more efficacious strategies.

The next section considers implications of the cognitive framework standard setting practice. Although presented in terms of practice, these ideas may also be investigated through either controlled experiments or field research. Some of the implications might already be part of

some standard setting practices. The hope is that a cognitive framework can both suggest new practices (or modifications of existing practices) and lead to a greater appreciation of practices that have theoretical basis in the cognition underlying standard setting panelist judgments.

Implications for Standard Setting Research and Practice

If we consider the cognitive framework as an accurate reflection of panelist cognition during the judgment task—the "kernel" of standard setting—then the practices surrounding this kernel should facilitate the outlined reasoning. This theory of probabilistic reasoning suggests that the standard setting meeting should be set up to aid panelists' understanding of the BE definition and items in terms of attributes (knowledge, skills) and their values (cognitive levels), and to cueing panelists' memory of student solution attempts. Below is a sampling of implications of the cognitive framework organized by the common elements of a standard setting meeting.

Recruit Panelists

The probability estimation framework and results from alternative models suggest a specific type of expertise needed by standard setting panelists. Many of the prescriptions in this section are consistent with prior writings on selection of standard setting panelists (e.g., Hambleton, Pitoniak, & Copella, 2012; Raymond & Reid, 2001) but add some theoretical justification based on the cognitive framework presented earlier.

Panelists should have direct experience with the population of examinees who take the test so that the panelists can draw on experiences of seeing these people respond to questions or tasks like those on the test. Such experience serves as a basis for judging the likelihood that a BE (a type of examinee) would respond to a given test item correctly. Panelists should also have sufficient expertise in the domain of the test to be able to discuss the skills needed by successful examinees and to be able to recognize the skills required to respond to a test item correctly. Panelists should also have sufficient expertise to understand the different cognitive levels of skills—both how they might be demonstrated by examinees and as a way for understanding the characteristics of test items that affect difficulty.

As part of the documentation for a standard setting study, information confirming the above types of expertise should be collected. Many panelists are educators, so they should have direct experience with students' performance in the domain. On the other hand, the expertise

outlined above suggests that people without such direct experience might not have suitable background for estimating the likelihood of solution. For example, consider the *PRAXIS*® French assessment, a licensure test that is one element of being certified to teach French in middle schools or high schools in some U.S. states. The test-taking population is prospective French teachers. One might think that an experienced French teacher would have direct knowledge of his or her students' success at French test items and, therefore, might understand the skills required for the items and the necessary cognitive level of those skills. Such a panelist might also understand the skills required of a beginning French teacher (the BE definition), but likely only from his or her own experience. However, an experienced French teacher likely would not have seen a variety of prospective French *teachers* attempting various tasks in French, which is the experience that is needed for standard setting judgments, according to the cognitive framework. Instead, a teacher trainer, such as a faculty member in a university's school of education, would have such expertise and would be a more appropriate standard setting panelist.

Discuss the Test

The probability estimation framework suggests that test familiarization should focus on a cognitive analysis of test items. Panelists should discuss what makes some test items easier or more difficult, outline the knowledge and skills (e.g., algebra, knowledge of French idioms) needed for the items, and describe the cognitive level (e.g., basic algebra skills, knowledge of common French idioms) of the skills required to solve the item correctly. In other words, the group discussion should form the basis for the independent analyses that each panelist will do. Many standard setting guidelines discuss the danger of having the test familiarization session turn into a test critique (e.g., Zieky et al., 2008, p. 182). However, many guidelines do not directly discuss the specific goal of helping panelists to accurately assess the skills (and their level) that items demand. One exception is the instructions for the Bookmark method in which panelists explicitly discuss what makes an item more difficult than the immediately preceding item in the ordered booklet. Panelists have some expertise in this area by virtue of being selected to participate, and they will likely be able to focus on the critical problem-solving elements of items rather than being distracted by surface features irrelevant to the difficulty of an item (Chi, Feltovich, & Glaser, 1981). A possible line of research might investigate the impact of training and meeting facilitation strategies that emphasize this focus on panelists' analysis of test items.

Define the Borderline Examinee

The definition of the BE derives from the PLDs. While the PLDs provide an overview of the knowledge and skills that should be possessed by someone who achieves a particular level on the test, the BE definition describes the slice of those skills possessed by the test taker who barely meets the requirements of the level. Thus, the PLD narrows the entire set of skills in a domain and the BE definition further narrows the domain, specifying the minimum necessary knowledge or skills needed to reach a level of performance.

Much has been written about the development of good performance level descriptions (e.g., Perie, 2008), but not as much concerning the distinction between the PLDs and the BE definition (although Egan, Schneider, & Ferrara, 2012, make this distinction in their discussion of Range vs. Target PLDs). However, we might assume that the good qualities of a PLD should be reflected in a BE definition. For example, best practices for PLD development (summarized by Perie, 2008) suggest that a PLD should consist of a clear label for each performance level (e.g., basic, proficient, advanced), a definition of that label that describes the general qualities of an individual reaching that level of performance, and a fleshed-out definition of the level that includes specific knowledge and skills.

According to the cognitive framework, when making a probability judgment (standard setting judgment), panelists compare the BE definition to their analysis of an item in terms of the attribute–value overlap between the two. This overlap might be considered in terms of the actual words in the BE definition, or the definition might be used as a cue to memories of actual student performance. In either case, the critical question is "What information should the BE definition contain to facilitate probability judgments?"

Primarily, the definition should include both the description of knowledge and skills as well as their cognitive level. This information corresponds to the attributes and values—the basis of probability judgment—in the cognitive framework. Interestingly, nowhere in Perie's (2008) review of PLD development, nor in other discussions of PLDs (e.g., Cizek & Bunch, 2007; Zieky et al., 2008), is the recommendation to include the cognitive level of the knowledge and skills. Egan et al. (2012) did highlight the need for cognitive complexity information in their discussion of target PLDs (corresponding to the BE definition) at a broad level, although they do not specifically recommend including such information within each statement of a target PLD. Yet not only the description of what knowledge and skills are possessed (attributes of the BE),

but also the cognitive level of said knowledge and skills (values of these attributes) are the key inputs to a cognitive model of probabilistic reasoning.

There are several ways that the cognitive level of knowledge and skills might be represented in a BE definition. The levels might be mentioned explicitly, as described in the examples earlier (e.g., knowledge of common French idioms; the examples of Egan et al., 2012, included just this type of information), or described through performance indicators. Performance indicators provide specific examples of what a BE can and cannot do. Such indicators may implicitly set a level of performance for the knowledge and skills described in the BE definition.

Additionally, research on decision making suggests that the properties of the information presentation affects the use of that information. For example, care should be taken with performance indicators and other methods of implicitly describing knowledge and skills, and their levels, in a BE definition. As mentioned earlier, people have a cognitive bias whereby mentioned information (i.e., concrete examples) may carry more weight in decisions than information only implicitly given (McCaffery & Baron, 2006). Although a BE definition cannot be comprehensive without becoming unwieldy (i.e., it is not feasible to state explicitly every knowledge and skill), standard setting facilitators should be aware that panelists might not incorporate into their judgments those same aspects of a BE that are only implied in the definition. Research might investigate approaches to training panelists that could help them avoid this bias (weighting explicitly given information more heavily) as well as other biases.

Train the Panelists

The previous discussion contains several recommendations on how the panelists should approach the individual judgment task (i.e., the "kernel" of standard setting), which has been described as a process of probabilistic judgment. These recommendations suggest some elements of panelist training that should occur. For example, as described in the Discuss the Test section, panelists should be encouraged (or taught) to conduct a type of cognitive analysis of the test items, which might facilitate their later comparison of the cognitive demands of the items to the knowledge and skills (and their level) in the BE definition. Additionally, panelists might be warned of the types of biases that can affect probabilistic estimations (e.g., the concreteness of the information, as mentioned earlier). The cognitive framework also suggests that the practice of helping panelists to think of specific people that match the BE definition may aid judgments.

This practice has occasionally been mentioned in the literature (e.g., Zieky et al., 2008), but without clear motivation. The practice might help panelist to apply their memory of student performance more consistently, leading to better outcomes for the standard setting meeting, such as greater interpanelist and intrapanelist agreement.

Ultimately, however, research is needed on the training and facilitation strategies that might best guide panelists to focus on the correct information, consider the correct aspects of their expertise, and avoid common biases during their standard setting judgments.

Conclusions

This paper outlined a cognitive framework for understanding the judgments that panelists make in a standard setting study. The framework outlined both the key elements that should influence the judgments—the test items and the definition of a BE—as well as the aspects of those elements that are important, namely the specification of knowledge and skills as well as their cognitive levels. The paper suggested two possible mechanisms whereby panelists reach a probability estimate based on this information. However, more research is needed to carefully specify a cognitive model that is tailored to the standard setting situation.

Much happens during a standard setting meeting, and it is difficult for a facilitator to know what aspects of the meeting will affect the reasoning of panelists. Zieky (2001) pointed out that standard setting research has brought a greater appreciation of how the nuances of conducting a standard setting study (e.g., instructions to panelists, training and practice on the methodology, the ways in which the study is described to panelists), rather than simply the methodology being implemented, have a profound influence on the outcomes of a study. A framework such as the one outlined in this paper provides a theoretical basis for (a) making decisions about how to conduct a standard setting to facilitate the cognitive processes of standard setting panelists, (b) guiding the decision of standard setting studies for new types of tests or otherwise situations that call for other than a traditional implementation of standard setting methodology, and (c) guiding research into the standard setting process.

References

Astebro, T., & Koehler, D. J. (2007). Calibration accuracy of a judgmental process that predicts the commercial success of new product ideas. *Journal of Behavioral Decision Making, 20,* 381–403. https://doi.org/10.1002/bdm.559

Bejar, I. I., Braun, H. I., & Tannenbaum R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and Modeling Cognitive Development in School* (pp. 1–30). Maple Grove, MN: JAM Press.

Brandon, P. R. (2004). Conclusions about frequently studies modified Angoff standard-setting topics. *Applied Measurement in Education, 17,* 59–88. https://doi.org/10.1207/s15324818ame1701_4

Bryant, D. J. (2007). Classifying simulated air threats with fast and frugal heuristics. *Journal of Behavioral Decision Making, 20,* 37–64. https://doi.org/10.1002/bdm.540

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152. https://doi.org/10.1207/s15516709cog0502_2

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30,* 93–106. https://doi.org/10.1111/j.1745-3984.1993.tb01068.x

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

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: MIT Press.

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard-setting study. *Applied Measurement in Education, 18,* 257–267. https://doi.org/10.1207/s15324818ame1803_4

Friedrich, J., Lucas, G., & Hodell, E. (2005). Proportional reasoning, framing effects, and affirmative action: Is six of one really half a dozen of another in university admissions?

Organizational Behavior and Human Decision Processes, 98, 195–215. https://doi.org/10.1016/j.obhdp.2005.06.002

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes, 103,* 277–290. https://doi.org/10.1016/j.obhdp.2007.01.003

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and Innovations* (2nd ed., pp. 47–79). New York, NY: Routledge.

Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). *Anchor-based methods for judgmentally estimating item difficulty parameters* (LSAC Research Report Series No. LSAC-CTR-98-05). Newtown, PA: Law School Admission Council.

Hein, S. F., & Skaggs, G. E. (2009). A qualitative investigation of panelists' experiences of standard setting using two variations of the bookmark method. *Applied Measurement in Education, 22,* 207–228. https://doi.org/10.1080/08957340902983978

Hein, S. F., & Skaggs, G. E. (2010). Conceptualizing the classroom of target students: A qualitative investigation of panelists' experiences during standard setting. *Educational Measurement: Issues and Practice, 29*, 36–44. https://doi.org/10.1111/j.1745-3992.2010.00174.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Karelaia, N. (2006). Thirst for confirmation in multi-attribute choice: Does search for consistency impair decision performance? *Organizational Behavior and Human Decision Processes, 100,* 128–143. https://doi.org/10.1016/j.obhdp.2005.09.003

McCaffery, E. J., & Baron, J. (2006). Isolation effects and the neglect of indirect effects of fiscal policies. *Journal of Behavioral Decision Making, 19,* 289–302. https://doi.org/10.1002/bdm.525

McGinty, D. (2005). Illuminating the "black box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education, 18,* 269–287. https://doi.org/10.1207/s15324818ame1803_5

Moon, J., Finn, B., LaMar, M., & Katz, I. R. (2018). Simulations of thought: The role of computational cognitive models in assessment. *ETS R & D Connections* (No. 26). Princeton, NJ: Educational Testing Service.

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 600–620. https://doi.org/10.1037/0278-7393.31.4.600

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing, 27,* 261–282. https://doi.org/10.1177/0265532209349472

Papageorgiou, S., & Tannenbaum, R. J. (2016). Structuring standard setting within argument-based validity. *Language Assessment Quarterly, 13,* 109–123. https://doi.org/10.1080/15434303.2016.1149857

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues & Practice* (winter), 15–29. https://doi.org/10.1111/j.1745-3992.2008.00135.x

Raymond, M. R., & Reid, J. R. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Lawrence Erlbaum.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 18,* 233–256. https://doi.org/10.1207/s15324818ame1803_3

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F.Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General, 137,* 73–96. https://doi.org/10.1037/0096-3445.137.1.73

Zhang, J., Hsee, C. K., & Xiao, Z. X. (2006). The majority rule in individual decision making. *Organizational Behavior and Human Decision Processes, 99,* 102–111. https://doi.org/10.1016/j.obhdp.2005.06.004

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51). Mahwah, NJ: Lawrence Erlbaum.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.