A System of Performance Standards:

Policy Reviews as an Additional Layer of Validity

Juan M. D'Brot

National Center for the Improvement of Educational Assessment

(Center for Assessment)

Paper presented at the annual meeting of the

National Council on Measurement in Education, Washington, DC.

Correspondence concerning this paper should be addressed to

Juan M. D'Brot
Senior Associate
Center for Assessment

jdbrot@nciea.org

A System of Performance Standards: Policy Reviews as an Additional Layer of Validity

The K-12 education sector has been flooded with change over the past decade. With states embracing the use of criterion-referenced tests as a result of the passage of the *No Child Left Behind Act of 2001,* the entry and exit of states in and out of assessment consortia, requests for Flexibility from the *Elementary and Secondary Education Act* (ESEA), and most recently the passage of the *Every Student Succeeds Act* (ESSA), the expectations set forth in state standards and large-scale assessments have been critical—but rarely static—foundations for assessment and accountability systems. Couple these factors with state-specific contexts around legislation, graduation requirements, and intensifying debates regarding local control, large-scale assessments are increasingly applied in high-stakes decisions.

As a result of these changes, there has been an increased departure from performance standards and cut scores simply reflecting the knowledge and skills acquired by students in a given domain. That is, policy and law is stretching the interpretation of cut scores, especially in the aggregate, far beyond claims regarding student learning (e.g., to make claims of educator effectiveness, to make decisions about student promotion/graduation, etc.). By establishing a comprehensive plan that includes procedural aspects aimed to directly incorporate consequential considerations, practitioners and policy makers may be better equipped to apply recommendations and establish clear communication points with diminished risk of deteriorating the assessment system. This can be done by explicitly embedding a policy review into the assessment development process. For the purposes of this paper, a policy review is defined as the process used by a state or a governing agency to examine the reasonableness of a *set* of recommendations following a standard setting that considers the coherence, consistency, viability, and consequences of an assessment's cut scores.

Increased Risk in an Age of Accountability

It is not surprising that states face an increased risk with regard to stretching the scope and application of large-scale assessment results. While not universally applicable, the turmoil states are facing has accelerated rapidly since 2010 with regard to standards, assessments, and their use. Several visible events have contributed to what seems like a recurring backlash against standards, large-scale assessments, and an expansion of how they are applied. These include, but are not limited to

- The release of Race to the Top in 2009;
- Widespread adoption of the Common Core State Standards in 2010 (i.e., at one point, 45 states, the District of Columbia, four territories, and the Department of Defense Education Activity adopted the Common Core);
- States being offered flexibility from ESEA in 2011, with 45 states, the District of Columbia, Puerto Rico, and the Bureau of Indian Education submitting requests; and
- The formation and operationalization of various assessment consortium, which includes Smarter Balanced, the Partnership for the Assessment of Readiness in College and Careers (PARCC), National Center and State Collaborative Partnership (NCSC), Dynamic Learning Maps Alternate Assessment Consortia (DLM), Assessment Services Supporting English Language Learners Through Technology Systems Consortium (ASSETS), and the ELPA21 Consortium.

These often charged incidents have contributed to the politicization of high-stakes assessment and have led to increased scrutiny over how assessment results and student performance are applied as consequences. These activities have been considered by some to be evidence of federal overreach and even as covert attempts to coerce states to adopt the current

administration's educational priorities—this is despite the fact that the majority of these efforts have been state-led (e.g., CCSS). However, a very real set of consequences is that assessment systems have been tied directly to educator effectiveness and the use of aggregate status and growth scores have been applied to school classifications that are arguably more comprehensive than those established under NCLB.

This increased scrutiny has placed high-stakes assessments and the standards on which they are based even more firmly in public discourse and as the target of state legislation. While an unfortunate consequence may be the need to quickly transition from one standards and assessment system to another, a more manageable consequence may be the stretching of the intended uses of an assessment. For example, many states are tasking their assessment or assessment system to be applied beyond its developed intent, asking it to simultaneously inform educators on student progress, principals on classroom effectiveness, local administrators on school quality, and function as a gatekeeper for promotion decisions. While standard setting must take into account the uses of the assessment, the methodology alone obviously cannot mitigate the contextual issues surrounding assessment use. However, incorporating certain steps in planning the standard setting process may serve multiple needs through thoughtful design—specifically, this paper makes the case for the incorporation of a policy review following the standard setting to mitigate a variety of issues that occur throughout the assessment development process.

Standard Setting to Define Cut Scores

Standard setting, which blends psychometrics, policy, art, and science (Cizek, 2001) relies on a comprehensive planning process that defines the theory, sequence, and structure through which participants will make recommendations. While various methods exist, standard

setting approaches commonly rely on high-quality data and "systematic, reproducible, objective, and defensible" information that varies depending on the on the format of the test, the types of information available, and context of the test (Cizek, 2012, p. 8).

Hambleton, Jaeger, Place, and Mills (2000) provide a popular classification scheme that delineates standard setting approaches into (a) judgments on a review of assessment material (e.g., Angoff, 1971; Bookmark Standard Setting Procedure, BSSP, Lewis, Mitzel, Mercado & Schulz, 2012), (b) judgments on examinee work (e.g., Body of Work, Kingston & Tiemann, 2012; Jaeger & Mills, 2001), (c) judgments on score profiles (e.g., Policy-Capturing, Plake, Hambleton, & Jaeger, 1997), and (d) judgments on examinees (e.g., Borderline and Contrasting Groups, Jaeger, 1989). Despite these various methodological approaches, there are several key steps that will likely yield reliable results in setting performance standards. These include the appropriate selection of a method, choosing and training a large and representative panel, collecting ratings and providing feedback, and collecting sufficient evidence of validity that includes evaluations, recommendations, and technical documentation (see Hambleton, Pitoniak, & Coppella, 2012).

It can be argued that goal of the process, regardless of the method, is defensibility of the conclusions and replicability of the process. Defensibility, however, depends in part on the context in which those performance standards, and subsequently cut scores, will be used. It may be the larger context that poses potentially the greatest challenge in setting performance standards. The notion of consequential use and its impact on the validity of a test or process (Linn, 1997; Moss, 1998) has become a major factor in setting performance standards. Newton and Shaw (2015) note the challenge in whether one should define validity as both the measure and its use, or to treat those two constructs as separate. Regardless of one's ideological position,

this challenge exists in setting performance standards. While psychometricians and scholars are primarily responsible for the design and delivery of a sound standard setting, one may not be able to separate the process (i.e., akin to the measure) and the application of the cut scores (i.e., akin to its use). This duality underscores the notion that standard setting is where psychometrics and policy intersect most acutely.

That is not meant to imply that setting performance standards has not historically included the notion of consequences during design and delivery (Brown, 2012; Cizek, 2012). Selecting the appropriate methodology requires an understanding of the context in which the assessment's cut scores will be used. Often in standard setting, concerns associated with the use of the test scores emerge during participant interactions and recommendations. Subsequently, facilitators may work to focus the activity of making judgments as content-oriented in nature. Facilitators may address consequentially-related concerns by noting they will be taken into account outside of standard setting, often by policy makers or state staff.

This attempt to focus participants to the content-based task at hand has the risk of being dismissive and has the potential to influence participants' cognitive processes. The effect on the cognitive processes of panelists is of such great concern that methods have been developed to minimize cognitive demand (e.g., BSSP, Lewis, et al., 2012) and several suggestions have been made specific to problems around participant understanding, demonstrating that understanding, and the complications associated with impact data (see Skorupski, 2012). However, external concerns may pose lingering threats to the recommendations made by participants.

Tversky & Kahneman (1974) raise the idea of a cognitive bias known as the anchoring effect. The anchoring effect inhibits the degree to which an individual will differ from their initial judgment on a given stimulus. While this is often used in examples like starting salary or

cost of an item, it can be applied to the example of a discussion among participants involved in a standard setting. For example, a participant might initially believe that a student must answer 75% of the items correctly to meet the target of proficient. Participant two, three, and four may believe that a student must answer approximately 50% of the items correctly to meet the target of proficient. Despite the greater amount of support for the 50% target, it is unlikely that the first participant will be swayed to adjust their recommendation to 50%. However, there may be a greater likelihood that their recommendation may adjust to a 70% or 65%.

While the goal of discussion among participants is to promote perspective sharing and mutual understanding of the content, PLDs, items in question, and process, there is a degree of consensus building that occurs. Interestingly, participants will often seek to understand an outlier's perspective, but will also engage in one of two strategies: the door-in-the-face or foot-in the-door technique to promote compliance to the group. The door-in-the-face (Cialdini, et al., 1975) is employed when a large adjustment is denied, but a smaller, more reasonable adjustment is made. Conversely, the foot-in-the-door technique (Freedman & Fraser, 1966) is used when a very small adjustment is made and is then followed up with a larger adjustment. Both approaches are often employed around those "areas of uncertainty" that may emerge among groups of items, samples of student work, or profiles of students.

These negotiations, however, rely on group dynamics being unfettered by outside influences. In the case of heightened accountability and consequences, cognitive anchoring may play a greater role due to outside factors. That is, an anchoring bias can manifest itself based on contextual issues or concerns thereby limiting natural adjustments that occur through within-group and across-group discussions. For example, a participant who is predominantly concerned with the consequential outcomes of a standard setting may artificially lower his or her

recommendations and attempt to adjust the group's ratings, thus lowering the overall recommendation. This may stem from concerns with growth-to-standard targets (see Betebenner, 2009) that might be used for school accountability or aggregates of student performance in educator effectiveness calculations.

While an expert facilitator may be able to navigate such an issue, it may be more effective to embed a step in the process to function as a "parking lot" of sorts for consequential concerns. The inclusion of this step—a policy review—as part of the initial standard setting planning process can have significant benefits. While policy reviews are referenced in texts (e.g., Egan, Schneider, & Ferrara, 2011; O'Malley, Keng, & Miles, 2012) and used in some state programs (e.g., State of Texas Assessments of Academic Readiness, West Virginia Education Standards Test), the relative frequency with which they are mentioned and applied in state programs is apparently far less when compared to the number of standard settings conducted.

The addition of a policy review can lead to benefits before, during, and after the standard setting. Before, it can ensure that the standard setting is designed in such a way that upholds the theory of action of the assessment system and its role in the state's educational system. During the standard setting, it can allow for participants to distance themselves cognitively and emotionally from the consequences associated with the cut scores and focus directly on the knowledge, skills, and abilities associated with the performance standard in question. After the standard setting, it can provide an additional layer of validity evidence and practical considerations for the implementation of recommended standards. These benefits are explicated in more detail in the remainder of this paper.

The Need for Explicitly Defined Policy Reviews

Standard setting participants make their recommendations by adhering to a process and by undergoing training in a methodology steeped in content (e.g., Angoff, Body of Work, BSSP). Panelists are asked to leverage their subject matter expertise in instruction, curriculum, standards to make recommendations about the knowledge, skills, and abilities that are expected of students in particular performance levels using content-based Performance Level Descriptors (PLDs, see Eagan, Schneider, & Ferrara, 2012). While there is the risk of a lack of consistency in cut scores across grades and contents, methods exist to promote consistency in cut scores.

Cizek and Agger (2012) discuss the need for, a review of, and approaches to implementing vertically moderated standard setting—a key tactic in developing defensibly articulated cut scores. In a well-designed standards and assessment system, inclusive of the standard setting effort, consistency can be promoted through the use of one or more of several methods (e.g., develop articulated standards, use articulated PLDs, use cross-grade panels, engage in cross grade-level discussions, host vertical articulation session, or calculate interpolated cuts) to support a coherent system of cut scores. This coherence benefits more than the assessment system itself and extends to the acceptance and use of the assessment. By leveraging a system of cuts that communicates a relatively steady set of signals to educators and administrators, states can help promote buy-in of the assessment system.

However, even when thoughtfully incorporating activities like those stated above during a standard setting event, recommended cut scores can vary significantly across grades and content areas. These inconsistencies subsequently raise concerns about the implementation of cuts that can emerge annually with the release of performance data. These inconsistencies are then exacerbated by the presence of external consequence factors. Further, states differ on the

degree to which participants are provided external benchmarks to which they can compare their own recommendations (see Phillips, 2010, 2012).

From the author's experience, the tone of standard settings for high-stakes assessments can range widely. In an attempt to separate consequences, states have encouraged standard setting participants to make judgments using only content-based rationales and minimizing the use of impact data. In other cases, states have incorporated external information (e.g., policy-based needs, benchmark data, historical state cuts compared to NAEP cuts) to inform (and in some cases persuade) participants to establish higher cut scores. These approaches essentially differ in whether cognitive anchors are defined by participants' own subject matter expertise or external signals of what kinds of knowledge, skills, and abilities are expected of students in a given domain. Both situations can benefit from the formal incorporation of policy review committees.

A policy review committee may provide additional evidence supporting a consistent, cohesive, and aligned set of cut scores for an assessment system that factor in context beyond content. The standard setting provides a set of strong content-based recommendation and forwards performance expectations against the knowledge and skills students should have at each performance level. The policy review committee would then provide a perspective that considers the outcomes and consequences of the recommended cut scores. This allows policy reviewers to step away from the content and focus on consistencies across all grades. Further, policy review committee members have the freedom to consider the benefits and risks associated with the implementation of recommended cut scores. Considering the current educational landscape with expectations for increasingly rigorous standards and PLDs, there are myriad

considerations around topics that include student and teacher goals, accountability, educator effectiveness, existing policy, policy development, and resource allocation.

Consider the example of a state whose focus is two-fold: (1) satisfying federal accountability by developing college and career ready indicators of student performance and (2) satisfying state-specific needs by establishing signals of minimum competency for promotion and graduation. These goals may exhibit some conceptual overlap but could be incongruous in practice. Let us say the first goal is informed by something like the recent Honesty Gap report (Achieve, 2015), which showcases the distance between NAEP's proficiency cuts and state-specific proficiency cuts. An equally likely scenario would be that the state has to demonstrate evidence of proficiency cuts that are reflective of college- and/or career-readiness to satisfy peer review requirements (Center for Assessment, 2015; U.S. Department of Education, 2015). In either of these cases, state policy-makers may highlight the disconnect between their historical state-specific cut score (e.g., an average of 70% proficient) and the NAEP-based signal (e.g., approximately 40% proficient).

Let us also say that the second goal would be to use the assessment for decisions associated with promotion and graduation. While initially seeming like one goal may be an extension of the other, the realities of requiring remediation or preventing a student from graduating unless they demonstrate mastery on a set of locally defined, similarly high-quality requirements becomes a daunting juggling act of resource allocation. If the state's only goal were to narrow this signal gap or to establish cuts based on promotion or graduation, a relatively straightforward conversation could be had. However, these goals become problematic considering a sample target of 40% proficient on the statewide assessment.

By no means does the inclusion of a policy review mitigate this very serious example. However, by explicitly including a policy review in the planning of a standard setting, psychometricians and practitioners can collaboratively plan for the types of issues and practical concerns that will arise during the life cycle of development and implementation. What the policy review will provide is then an opportunity for concretizing problems in advance, ensuring that the theory of action of an assessment system is in line with the types of claims being made in standard setting, and determining alternative approaches (e.g., determining whether additional cuts should be considered that communicate different signals for different purposes).

Policy Reviews as a tool for Planning and Evidence Gathering

A policy review can yield benefits before, during, and after standard setting by refocusing those involved on the purpose and meaning of the cut scores on an assessment, the intended uses of the cut scores, and the implementation of those cut scores. The following sections describe benefits of implementing a policy review before, during, and after the standard setting.

Before the Standard Setting

The design of an effective assessment system should begin with explicating a clear theory of action (see Marion, 2010). A theory of action should highlight the components of an assessment and how they are connected, including the "hypothesized mechanisms or processes for bringing about intended goals" (Marion, 2010, p. 1). When planning a standard setting, including a policy review prior to the actual standard setting process can serve as an important filter (and reminder) to help policy-makers think about the impact of cut scores. This is especially important when cut scores will be used for multiple signals (as in the previously described example). Further, this can facilitate advanced planning in policy-focused discussion about the use of the assessment cuts, how to establish those cuts (i.e., a CCR-focused cut and a

promotion-focused cut), or how to create compensatory rules (see Mehrens, 1990) to mitigate issues that are in conflict with the state's theory of action.

During the Standard Setting

As noted previously, standard setting participants engage in a social process that often involves some negotiation depending on the discrepancy of their recommendations. This cognitive process relies on the subject matter expertise of participants to make recommendations through an iterative process of introspection, recommending a cut, and discussion (see Skorupski, 2012). However, this process is equally as social and emotional as it is intellectual. Participants share their perspectives in groups and may demonstrate a nervous awareness of how the cut scores will be used. The degree of anxiety or discomfort exhibited by participants while making judgments can then have implications on the kinds of recommendations that are made.

Thus, the inherently judgmental process of standard setting relies on participants being able to access their own experience and expertise without being distracted (from a purely procedural standpoint) by unrelated distractions. Because of their lack of measurement expertise, participants must be trained in and learn a new process (i.e., the standard setting methodology) in order to make recommendations using their content expertise (Skorupski, 2012). Therefore, the fewer impediments participants have to making sound and justified judgments, the more participants' recommendations are based in content, which can contribute to the defensibility of the standard setting.

Typically, participants are provided with an overview of the assessment system, the standard setting method and process, and how recommendations will be taken forward to a governing body (e.g., a State Board of Education, Technical Advisory Committee, or external performance standards review committee). By highlighting the role of a policy review

committee during the actual standard setting itself, facilitators can separate the recommendations of standard setting participants from the consequences associated with the recommended cut scores. This can minimize the anxiety for panelists and lessen the psychological strain associated with making recommendations on high-stakes assessments.

By distancing the application of the cut scores and asking participants to focus solely on the knowledge, skills, and abilities set forth in the PLDs, more direct and substantive conversation may emerge. However, this assumes that a sufficient number of performance levels are developed and those performance levels (and associated PLDs) meet the needs stated in the theory of action used to develop the assessment (e.g., CCR-focused cuts, minimum competency cuts, or remediation-based cuts). Thoughtfully planning for the types of discussions and anticipating the questions that may be raised by participants can further serve to validate the appropriateness of the intended use of the cut scores. This in turn can influence subsequent policy discussion at the state.

After the Standard Setting

The benefits posed before and during the standard setting require that a policy review be implemented so that participants honor the process of the standard setting and feel comfort in making recommendations in support of the assessment system and its application. If facilitated well, a policy review will yield at least three benefits to the larger assessment development and implementation process. These benefits include the following:

1. Provide an additional set of recommendations divorced from the content-based rationales that privilege a cohesive set of cut scores communicating consistency within the assessment;

2. Forward considerations for the implementation of the recommended (adjusted or not) cut scores that speak directly to the consequences of the assessment's use; and
3. Identify communication and implementation needs reflecting the concerns of those represented on the policy review committee.

These three benefits serve to add an extra layer of validity in the process of setting standards to ensure a balanced set of perspectives are presented when implementing cut scores. In the same way that we strive to establish multiple sources of validity evidence for assessments, it is becoming evident that there is also a need for multiple types of evidence when setting performance standards, especially from a consequential standpoint. The Standards (AERA, APA, NCME, 2014) highlight the idea that validity is an argument stemming from sources of validity evidence (i. e., based on test content, response processes, internal structure, and other variables). One can draw a similar parallel to the types of validity evidence that are garnered during standard setting. A standard setting privileges a content-focused view of evidence. A policy review, however, privileges a broader view of validity evidence by prioritizing the application of cut scores into a system of assessments. Used in conjunction, practitioners may be better equipped to support the implementation of the assessment and cut scores.

Another way to conceptualize this benefit is through one of balance stemming from a two-pronged approach of a standard setting and a policy review. The standard setting event is supported by participants' subject matter expertise in content, curriculum, and instruction. It is grounded in the standards, PLDs, and performance expectations for students. Standard setting focuses on the knowledge, skills, and abilities all students *should* demonstrate at each performance level (Hambleton, Pitnoiak, & Copella, 2012). The post-standard setting policy review, however, is supported by a wider view of the standards and assessment system. It is

grounded in the consequential considerations and consistency of the signals communicated by the assessment. A policy review focuses on the intended effects on instruction, accountability, and policy. While a standard setting does not require a policy review to be successful, a policy review's strengths can be borrowed early on in the assessment development process by maintaining a strong focus on the intended uses of the cut scores and assessment.

Planning for the Policy Review

Based on this author's experience, policy reviews have been implemented under two main conditions. In one case, they have been implemented as a stop-gap between a standard setting and presenting recommendations to a deciding body due to erratic recommendations or policy goals not being reached. On the other, they have been applied because policy makers and practitioners were aware of the difficulty associated with the implementation of policy-influenced cut scores (i.e., NAEP-like cuts on a statewide assessment). Ideally, a policy review would be implemented to arrive at a final set of recommendations for the entire assessment system and potentially establish a phase-in plan for implementation (see O'Malley, Keng, & Miles, 2012)

O'Malley and colleagues (2012) refer to the case of the State of Texas Assessments of Academic Readiness (STAAR) end-of-course (EOC) assessments with the purpose and expected outcomes expected from a policy review meeting. However, the way in which a policy review is planned, framed to participants, and implemented is of vital importance to its success. The subsequent section describes a potential approach to planning and executing a policy review meeting. It details the overview of the standards assessment system and its purpose, the role of the policy review committee, the role of PLDs, the standard setting methodology used, the results from the standard setting, an in-depth discussion of the consequences and context of the

state, capturing recommendations from the policy reviewers, and a discussion of communication and implementation considerations. These can be divided into four main portions of a policy review: Before, Front End, Main Discussion, and a Wrap-up of a policy review.

Before the Policy Review

Representation is critical, and does not differ much from the approach in identifying representative participants for a standard setting. However, the subject matter expertise is different from a standard setting. The primary concern during a policy review is about the use, communication, and resources associated with the cut scores being implemented. Thus, representatives should be limited to a number that would be manageable as a single group that still represents a diverse group of perspectives (i.e., 8-12 participants). Participants should include educational administrators (e.g., school principals, local education agency staff), governing body representatives (e.g., State Board of Education members, legislative representatives), higher education representatives (e.g., higher education agency members, post-secondary institution representatives, admissions officers), and an appropriate number of participants from the standard setting (e.g., 1 per domain or grade span). Further, the goal of the policy review committee should be vetted with the governing agency that will be approving the cut scores and the message should be succinct enough to be delivered in a succinct "elevator speech."

Front-end of a Policy Review

During the front-end discussion of a policy review, one would describe the overview of the standards assessment system and its purpose, the role of the policy review committee, the role of PLDs, the standard setting methodology used, and showcase the results from the standard setting. Participants would receive information upon arrival containing much of the information

typically provided during a standard setting (e.g., confidentiality agreement, travel voucher, demographic survey emergency contact form, agendas, etc.) to be completed before the session begins.

Following a welcome and introduction, the facilitator and the hosting agency (e.g., a state education agency) would provide the orientation and purpose of the meeting. This is an opportunity for the hosting agency to provide a detailed discussion on policy, the role of standards and assessments in the educational system, how the standards and assessment system has evolved over time, and the goals of the system.

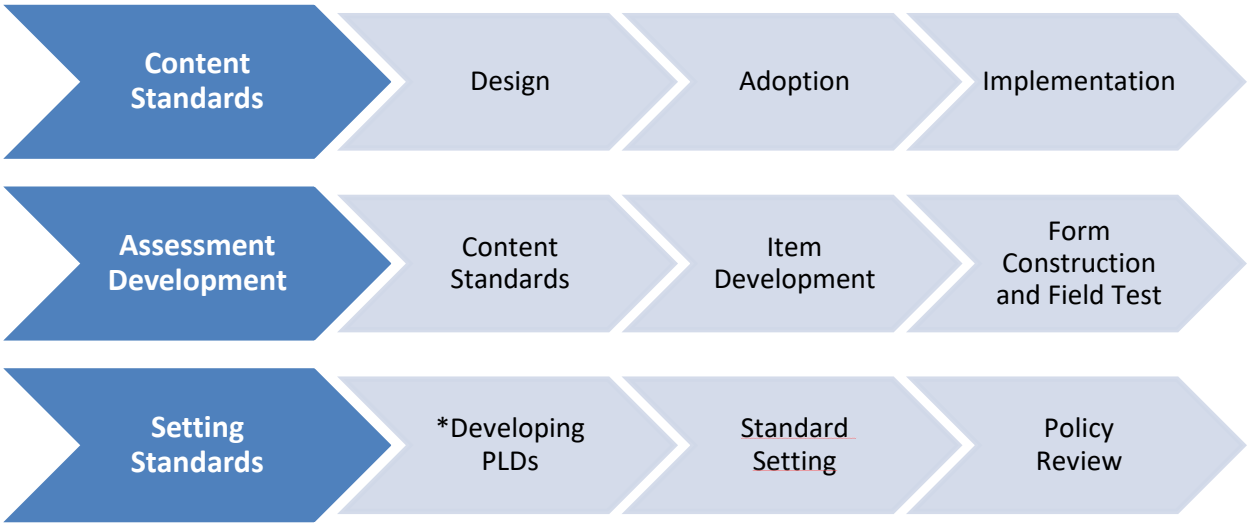
This can then segue into the facilitator describing the role of the policy review committee by using the succinct message that has been vetted with key stakeholders. While a majority of the meeting would be devoted to the main discussion around policy review recommendations, a critical portion of the meeting should focus on establishing the credibility of the standard setting methodology used. The facilitator should describe the standard setting method with sufficient detail that policy review participants have a high-level picture of what occurred during the standard setting. This also affords the representatives from the standard setting the opportunity to provide first-hand commentary or answer any questions posed by the remaining policy review committee members. It may be beneficial to provide meeting participants with samples of the materials used by the standard setting participants, with particular emphasis placed on the PLDs to highlight any relevant similarities or shifts in expectations from the previous assessment system.

The front-end of the policy review would then conclude with a review of the results from the standard setting. This review should include, minimally, summaries of the participants' evaluations (see Cizek, 2012) and impact data from the pre- and post-articulation results (see

Reckase and Chen, 2012). As previously noted with regard to the anchor bias, the way in which the standard setting is presented to the policy review committee is important so that participants do not dismiss the value or the importance of standard setting. While the policy review committee should have the freedom to discuss changes to the cut scores, a complete disregard to the standard setting recommendations would be difficult to defend as part of the larger development process. Establishing value and credibility in the process and participant recommendations is a valuable investment to ensure subsequent conversations are productive.

Main Discussion of a Policy Review

A majority of the policy review meeting will focus on an in-depth discussion of the consequences and context of the state, capturing recommendations from the policy reviewers, and arriving at a single set of recommendations to take to the governing body for review. It is helpful to reiterate the goal of the policy review committee and to frame the policy review as simply another step in the process of assessment development. One possible way to frame this to participants is by conceptualizing the assessment system development process as the following set of three interrelated paths, with setting standards concluding with a policy review:



*Clarification will likely be necessary noting that various types of PLDs are used throughout the assessment development process (see Egan, Schneider, & Ferrara, 2012).

Figure 1. Three inter-related paths of assessment development.

Time spent discussing the many stages in the process can be framed as checks and balances in assessment development to ensure evidence of quality is being collected over time. By reiterating the compounded information that is collected over the course of an assessment's life cycle, one may be able to better highlight the value of the conversations that take place during both the standard setting and the policy review.

The reiteration of the charge and review of the larger assessment development process can then transition into the consequential nature of cut scores. It is important to allow ample discussion space allowing policy review members to raise issues associated with how the cut scores will be used. While conversations can easily become circular, a skilled facilitator should strive to achieve saturation (i.e., an adequately representative number of issues relevant to the state) of major themes that emerge (see Mason, 2010). While far from exhaustive, common themes that emerge include student accountability, the impact of motivation and/or social promotion, educator evaluation, school accountability, incentivizing teaching and learning,

disincentivizing progress, demonizing the assessment system, and resource allocation (or lack thereof).

As participants transition into making recommendations, it will be up to the hosting agency to determine how much uncertainty they are willing to tolerate. For example, some states have set expectations that adjustments can only be applied if there is 100% agreement, while others have stipulated a super majority of greater than 2/3rds agreement is sufficient. Further, one will need to decide on the structure through which recommendations are made (e.g., Robert's Rules of Order to motion for a recommendation). Regardless of the criterion and process through which recommendations are made, facilitators and practitioners will also need to determine the extent to which "guard rails" will be used as policy review members make recommendations. This same concern is present with vertical articulation sessions during a standard setting where panelists can use standard errors to make adjustments to cut scores (Cizek & Agger, 2012; Lewis, 2001). A major difference, however, would be that vertical articulation often requires a secondary filter of using prior participant recommendations whereas a policy review may simply rely on the impact data associated with the cut scores.

The main discussion of the policy review would then conclude with the facilitator and any additional members of the facilitating team presenting the final set of recommendations to the policy review committee. Following a verification of the accuracy of the recommendations, the policy review committee can transition to the conclusion of the meeting.

Wrap-up of a Policy Review

The final portion of the policy review will be dominated by pragmatism and logistics. This section of the policy review should revolve around considerations for communication and implementation. It is important to remember that at this point, participants will have been

engaged in intense conversations and may be anxious to conclude the meeting. Additionally, the conversation throughout the meeting will likely have been peppered with ideas related to communication and implementation. Thus, it may be valuable to have an additional facilitator who can function as a note taker or to work with the hosting agency to verify policy or historical data. The additional facilitator may also be well positioned to track previous suggestions relevant to this portion of the policy review meeting.

Policy review participants should be asked about the real-world challenges they will face when the cut scores are implemented based on the final recommendations. This can serve as a springboard for recommending concrete actions and suggestions for outreach efforts, communication plans, and marketing materials. The discussion could also focus on the how the cut scores should be implemented. This might include the timeline for implementation, how to phase in the recommended cut scores, or ways in which to support policy or legislative mandates. Following these discussions, facilitators can provide participants with the meeting evaluation and adjourn the meeting.

Conclusion

The quality of an assessment rests on the inferences one can make from the evidence collected. While only a part of the assessment development process, establishing defensible cut scores relies on high quality processes and evidence. In an age of heightened accountability where assessments are often stretched beyond their intent, incorporating a policy review into the assessment development process supports a consistent focus on how assessment results may be used. This extends beyond the policy review if a plan is established early. The benefits of a policy review can be realized before, during, and after a standard setting. Thus, a policy review affords practitioners an additional layer of validity when setting standards.

Practical Considerations

As argued in this paper, the policy review represents a critical step in the assessment development process that has the potential to course-correct decisions related to policy throughout the assessment development process. Further, it provides a host of benefits when considering how to implement recommended cut scores. The following considerations are forwarded as ways to benefit from the explicit planning and use of a policy review.

1. Establish a theory of action that clearly articulates the guiding principles of an assessment system and includes the signals it is trying to communicate.
2. Establish sufficient buy-in for the assessment system and its downstream uses to insulate against change and misuse.
3. Consider holistically the applications of the assessment system and factor that into setting performance standards.
4. Embed explicitly vertical articulation and a policy review as an additional layer of validity evidence for the recommended performance standards.
5. Utilize the policy review committee as a source of developing a communication campaign that includes areas of concern and mitigation strategies.
6. Be mindful of the policy review serving to lower the standards in light of potential consequences.

References

Achieve (2015). *Proficient vs. prepared: Disparities between state tests and the 2013 National Assessment of Educational Progress*. Washington, DC: Achieve. http://www.achieve.org/publications/proficient-vs-prepared-disparities-between-state-tests-and-2013-national-assessment

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.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*, 42–51.

Brown, W. J. (2012). Moving forward: Educational, social, and population considerations in setting standards. In G. Cizek (Ed.), *Setting performance standards: Theory and applications* (2nd ed., pp. 571-579). New York: Routledge.

Center for Assessment (2015). *Annotated assessment peer review guidance*. Washington, DC: Council of Chief State School Officers.

Cialdini, R. B., Vincent, J. E., Lewis, S. K., Catalan, J., Wheeler, D., & Darby, B. L. (1975). Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of personality and Social Psychology*, 31(2), 206.

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

Cizek, G. J. (2012). The forms and functions of evaluations in the standard setting process. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp 165-178). Routledge: New York, NY.

Cizek, G. J. & Agger, C. A. (2012). Vertically moderated standard setting. In G. Cizek (Ed.), *Setting performance standards: Theory and applications* (2nd ed., pp. 467-484). New York: Routledge.

Egan, K. L, Schneider, M. C., & Ferrara, S. (2011). The 6D framework: A validity framework for defining proficient performance and setting cut scores for accessible tests. In S. N. Elliot, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 275-294). New York: Springer.

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

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: the foot-in-the-door technique. *Journal of personality and social psychology,* 4(2), 195.

Hambleton, R. K., Jaeger, R. M., Place, B. S., & Mills, C. M. (2000). Setting performance standards on complex educational assessments. *Applied psychological Measurement, 24,* 355-366.

Hambleton, R. K., Pitoniak, M. J., & Coppella, J. (2012). Setting performance standards. In G. Cizek (Ed.), *Setting performance standards: Theory and applications* (2nd ed., pp. 47-76). New York: Routledge.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Education measurement* (3rd ed., pp. 485-514). New York, NY: Macmillan.

Jaeger, R. M. & Mills, C. M. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives* (pp. 313-338). Mahwah, NJ: Earlbaum.

Kingston, N. M. & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The body of work method. In G. Cizek (Ed.), *Setting performance standards: Theory and applications* (2nd ed., pp. 201-223). New York: Routledge.

Lewis, D. M. (2001). *Standard setting challenges to state assessments: Synthesis, consistency, balance, comparability.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–253). New York, NY: Routledge.

van der Linden, W. J. (1995). A conceptual analysis of standard setting in large-scale assessments. In Proceedings of Joint Conference on Standard Setting for large-Scale Assessments. L. Crocker & M. Zieky (Eds.). National Assessment Governing Board: Washington, D.C. (pp. 95-115)

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*, 14–16.

Marion, S. F. (2010). *Developing a theory of action: A foundation of the NIA response*. Dover, NH: Center for Assessment.

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews. *Forum: Qualitative Social Research, 11,* 1-13.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17,* 6–12.

O'Malley, K., Keng, L., & Miles, J. (2012). From Z to A: Using validity evidence to set performance standards. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 301–322). New York, NY: Routledge.

Phillips, G. W. (2010). *International benchmarking state education performance standards.* Washington, DC: American Institutes for Research.

Phillips, G. W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 323–346). New York, NY: Routledge.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement, 57,* 400-411.

Reckase, M. D. & Chen, J. (2012). The role, format, and impact of feedback to standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 149–164). New York, NY: Routledge.

Skorupski, W. P. (2012). Understanding the cognitive processes of standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 135–147). New York, NY: Routledge.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases *Science, 185*, 1124-1131.

U.S. Department of Education (2015). U*.S. Department of Education peer review of state assessment systems: Non-regulatory guidance for states for meeting requirements of the Elementary and Secondary Education Act of 1965, as amended*. Washington, DC: ED. https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf