

Considerations for Establishing Performance Standards for Educator Evaluation Systems

Erika Hall, Elena Diaz-Bilello, and Scott Marion

National Center for the Improvement of Educational Assessment

6.2.2015

Introduction

Despite decades of district experience implementing educator evaluation systems (EES)¹, these systems have been criticized for failing to adequately hold educators accountable for student performance, provide appropriate incentives to improve teacher instruction and effectively differentiate teachers' performance (Burling, 2012; Weisberg, Sexton, Mulhern, & Keeling, 2009). Weisberg et al., argue that most pre-reform EES are flawed because personnel evaluations are factored into remediation and dismissal decisions, but typically not used to inform other critical teaching areas, such as professional development. Specifically, in most states and districts, performance appraisals for teachers are not designed to facilitate instructional improvement; instead, the underlying rationale for evaluating performance is punitive.

Due largely to waiver requirements submitted for the Elementary and Secondary and Educational Act (ESEA) or federal Race to the Top (RTT) guidelines, several states and districts have implemented or are piloting next generation EES. In an attempt to address criticisms of the old evaluation systems, many of these new EES encourage targeted feedback and data to help improve instructional practices and ultimately, student achievement (Minnici & Leo, 2013). However, despite the desire to provide teachers with formative feedback about instruction, Firestone (2014) notes that most EES are designed to meet administrative goals (i.e., sorting and identifying which teachers require probation and ultimately termination) rather than goals related to professional improvement. This fact is exemplified, in part, by a lack of clearly defined, coherent performance criteria for most EES.

In this paper, performance criteria will be defined in terms of two elements: qualitative statements that articulate the expectations associated with varying levels of performance in a given content area or domain (i.e., performance level descriptors) and the scores or ratings necessary to achieve each performance level given a specified assessment or tool (i.e., performance standards). Up to this point little attention has been paid to translating ratings and scores from EES into meaningful performance expectations for teachers. Instead, most researchers focus on the technical quality of component measures and the manner in which those measures should be combined to support decisions related to overall effectiveness (see Herman et al., 2011; Mihaly et al., 2013; Hansen, et al., 2013; Glazerman et al., 2011). While these technical and methodological issues are important to consider, they do little to support the direct use and interpretation of EES results.

¹ Throughout this paper, EES is used in both singular and plural form.

In this paper, we provide guidelines for establishing performance criteria within an EES. The document was written to support state and district practitioners— those typically tasked with designing such systems — but will serve as a useful resource for anyone interested in the development, review and evaluation of these and other personnel-based evaluation systems. More specifically, this paper focuses on the key principles and design considerations guiding the work to set performance standards. Although we touch upon common standard setting approaches, this paper does not provide step-by-step guidelines for conducting the standard-setting work. Procedures must be defined in consideration of key contextual factors and will therefore vary across sites. Instead, we highlight the importance of weighing key contextual factors when designing a standard setting approach. Readers seeking information on the various technical approaches to standard-setting should refer to the vast array of literature available on this topic (see for example, Cizek, 1996; Hambleton, et al., 2000; Kane, 1994; Shepard et al., 1993).

Standard Setting – A brief Introduction

The purpose of standard setting is to define and operationalize the expectations associated with different levels of performance on a specified measure of interest (Cizek, & Bunch, 2007; Hambleton & Pitoniak, 2006). Standard setting is largely a policy-making activity and, therefore, requires the input of different stakeholders at various stages of the process. For example, before standard setting can occur, decisions must be made about the number of performance levels needed, how they should be labeled, and the manner by which they will differentiate performance (Perie, 2008; Egan et al., 2011; Schneider et al., 2010). For the most part, these decisions are made by policymakers to communicate key goals to those using or interpreting the results.

Although a variety of standard setting procedures exist, common to all is the need for well-defined performance level descriptors. Performance level descriptors (PLDs) are explanations or summaries of the knowledge, skills, competencies, or behaviors associated with performance at a given level as defined by subject matter experts (Perie, 2008; Lewis & Green, 1997). Within the context of educational assessment, for example, PLDs describe the skills and abilities that a typical student within a given performance level would be expected to display given the type and range of skills assessed by the test. PLDs are intended to provide a clear, common definition of what it means to be classified within a given level and therefore must be understood and articulated well before performance standards (i.e., cut scores) are set (Perie, 2008; Bejar et al., 2007; Egan et al., 2011; Ferrara et al., 2009).

Establishing performance standards involves translating the competencies defined by the PLDs to the score scale metric so that valid score-based inferences can be made about what one knows and can do (Lewis et al., 1996). To establish performance standards for most K-12 assessments, panels of content experts review the PLDs in conjunction with test items, student responses, performance data and other materials relevant to the process, and engage in activities that result in cut score recommendations. In this context, cut-scores are defined in light of one performance measure (i.e., test score) and the range of ability in the test taking population is relatively well understood.

The process of establishing PLDs and performance standards becomes much more complex within the context of educator evaluation. This is due in large part to the use of multiple measures and complex data aggregation, in conjunction with a variety of other factors unique to this context. To clarify the nature of performance standards in EES, the section that follows describes two common but

competing purposes for educator evaluation and introduces the types of measures typically associated with an EES. Subsequently, we outline considerations related to the development of performance level descriptors and highlight key contextual factors that may influence the standard setting approaches ultimately applied. Finally, a hypothetical scenario reflecting a typical state or district EES is provided as an attachment to this paper to help illustrate how contextual factors influence the standard setting process.

To set the stage for this discussion, a summary graphic illustrating the relationship among key factors influencing the standard setting process for EES is provided in Figure 1, below. Each of these factors is discussed in turn within the context of this paper.

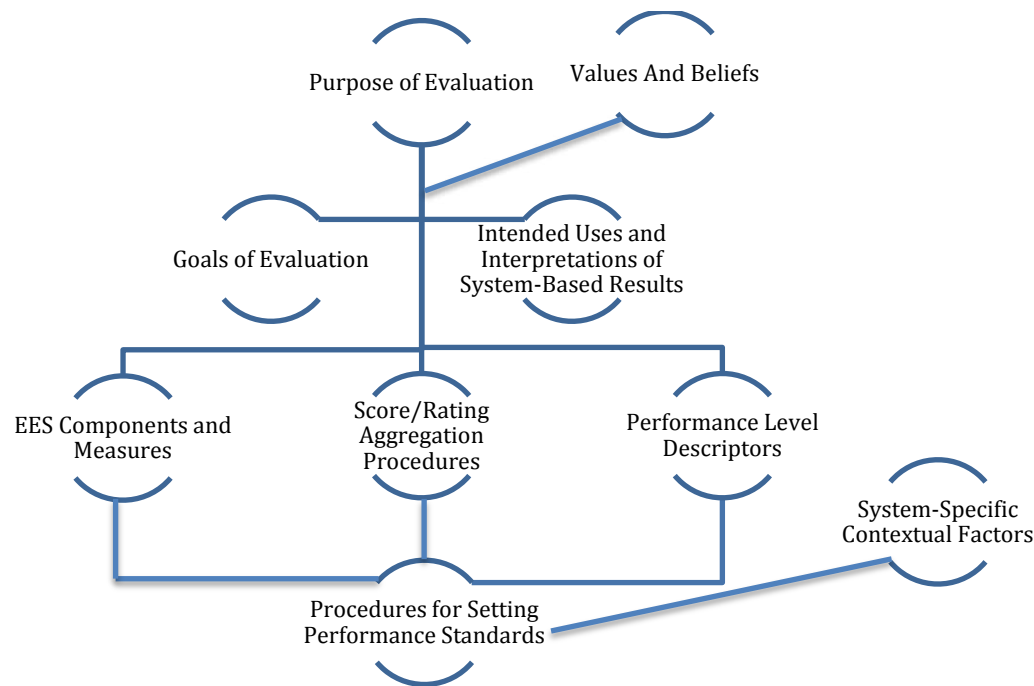


Figure 1. Elements Influencing the Development of Performance Standards for EES

Purposes of EES

The EES developed by many states and districts typically emphasize one of two purposes: to identify and remove poor performers, or to improve the workforce. These two EES purposes align with those identified in the Human Resources sector by Pulakos, (2004) and by the Society for Human Resource Management (SHRM) and the Approved American National Standard Institute (ANSI) in their Human Resources Performance Management Standard guide. According to Pulakos (2004) and SHRM and ANSI (2012), personnel evaluation is typically conducted to serve three purposes: to meet administrative or personnel management ends, to meet strategic ends, and to meet developmental or performance management ends. Table 1 below provides a brief description of each purpose, as well as a few examples of the types of goals associated with each.

Table 1. Different Purposes of Personnel Evaluations

Purpose	Description & Exemplar Goals
Administrative	<p>Support administrative decisions related to retention, promotion, selection</p> <p><i>Goals:</i></p> <ul style="list-style-type: none"> • identify and reward high performing employee • identify and remove low performing employees • increase employee motivation to improve through extrinsic rewards
Strategic	<p>Evaluate the extent to which there is alignment between an employee's goals and strengths and the work they are doing for the organization.</p> <p><i>Goals:</i></p> <ul style="list-style-type: none"> • improve employee satisfaction • increase efficiency and productivity • increase employee retention
Developmental	<p>Provide information and feedback to all employees to help them improve performance</p> <p><i>Goals:</i></p> <ul style="list-style-type: none"> • employee improvement – through provision of detailed feedback at employee level • organizational improvement - through identification of strengths/weakness at system level

According to SHRM and ANSI, a personnel evaluation system focused on administrative or personnel management lacks a connection to a broader systems theory of supporting human capital needs. Outcomes in an administrative system are not used to help develop inputs for the next evaluation cycle and rely largely upon salary rewards and bonuses to motivate employees (SHRM & ANSI). In contrast, the developmental purpose is designed to ensure that outcomes from one evaluation cycle feed into the next. This cycle connects this approach to a systems theory where feedback loops and reflections of current performance are used to define and align future performance objectives for employees (SHRM & ANSI).

Research suggests that the design of most EES systems is shaped by two competing motivational theories that correspond to the administrative and developmental purposes defined above (Firestone, 2012). Systems designed to inspire intrinsic motivation focus on using measures that establish feedback loops with teachers, whereas systems designed around extrinsic rewards are less likely to value measures linked to teacher improvement. According to Firestone (2014), it is very challenging to meet both purposes well and/or motivate the workforce using both intrinsic and extrinsic incentives, since each purpose requires different measures. He also notes that, for most RTT states, EES designs are largely dictated by a desire to motivate the workforce through extrinsic rewards or meet administrative goals.

Table 2 illustrates how the design of EES interacts with the underlying purpose for evaluation. Since most EES strive to serve *both* administrative and developmental purposes, one could imagine a middle column in Table 2 that reflects a scenario in which both purposes are equally weighted. In practice, the majority of EE systems prioritize one purpose over the other due to the difficulty of achieving both well (Firestone, 2014).

Table 2. Comparison of EES Design and System Relative to Intended Purpose

EES Design	Administrative	Developmental
Use of Results	Inform administrative decisions for pay increases, dismissals and remediation.	Facilitate employee improvement
Focus of Evaluation Efforts	Identify performance on one or more underlying scales to support sequencing, sorting and categorization of individuals to support decision making	Identify where there are gaps in performance relative to defined performance expectations so that they may be addressed
Stakes Associated with Performance	Moderate to High	Low to High
Frequency of Data Collection	As often as deemed necessary and appropriate to make reliable and accurate decisions.	As frequently as possible to support ongoing progress monitoring, evaluation and feedback loops with teachers.
Prioritized Data/ Information	Quantitative data points and objective outcomes.	Qualitative interpretations of quantitative data points that can be directly observed or based on measures closely aligned with instructional targets.
Outcomes Expected to Result from Evaluation	Data that supports aggregation and decision making in the most reliable and systematic manner possible.	Documentation of feedback that details and (possibly) quantifies gaps in employee performance relative to defined expectations in key areas specific to the job.
Major Concerns	Reliability, fairness, and objectivity of system-based measures and ratings.	Accuracy and completeness of information collected. Quality and validity of system-based recommendations and inferences related to improvement.
Role of performance expectations in generation of performance standards	Used primarily to support evaluation of the reasonableness of observed distributions of ratings or the specification of expected impact in light of defined expectations.	Necessary to support the specification of performance standards that inform improvement and plans for remediation.

Role of Stakeholders	Inform the specification of performance expectations. Review impact and procedures associated with proposed performance standards.	Key to the specification of performance expectations, as well as the manner in which data should be combined and reported to support improvement. Necessary to ensure performance standards align to defined expectations.
-----------------------------	--	--

Educator Evaluation System Design and Data Aggregation

Although state and district EES may vary greatly in form and function, most are comprised of measures that, on their own or in conjunction with other measures, serve to evaluate a teacher’s instructional practices and his or her contribution to student performance. Figure 2 presents a typical design for the educator evaluation systems currently specified across many states and districts to comply with ESEA waiver or RTT requirements. Under this common model, an overall effectiveness rating is comprised of data resulting from two larger *components (green)*: student outcomes and teaching practices. Under each component, different *measures (purple)* are scored and aggregated to generate a rating or a score at the component level. The component-level scores or ratings are then aggregated to provide an overall effectiveness determination (blue) for each teacher.

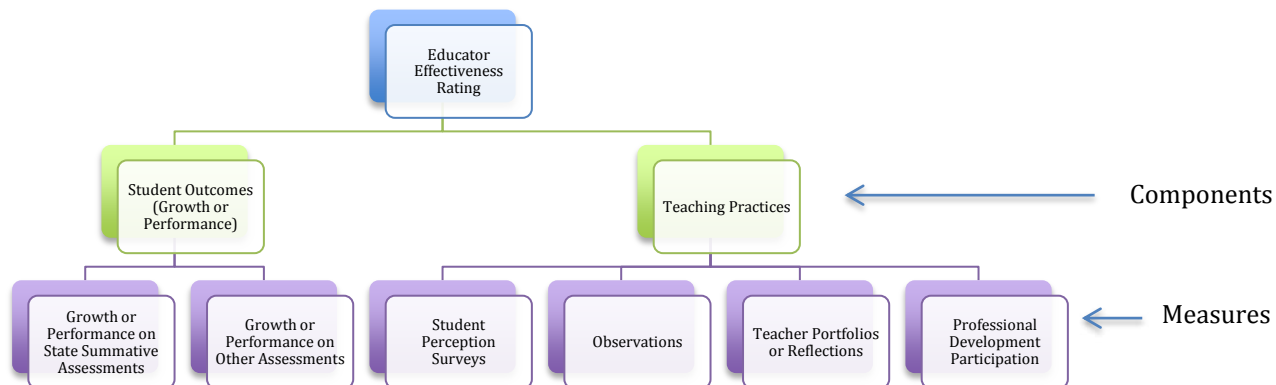


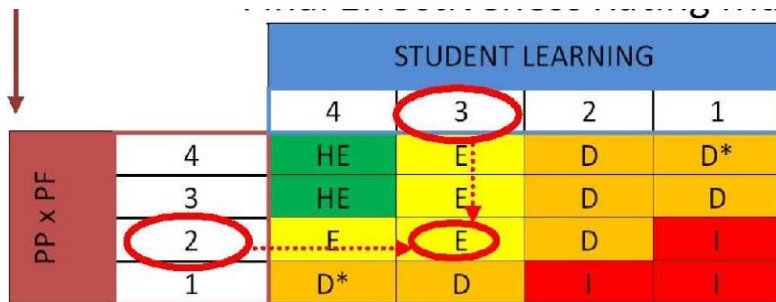
Figure 2. Common Elements Specified in an EES

The procedures used to aggregate data often vary by state and district; reflecting different beliefs as to how individual measures/components should be valued within the overall EES. Two approaches that serve to operationalize these beliefs are the conjunctive approach and the compensatory approach. A conjunctive approach specifies a set of conditions that must hold *across all measures* in order for a particular rating to be assigned. A rule such as “The teacher must obtain a Student Outcome rating of 3 *and* a Teaching Practices Rating of 4 in order to receive an overall rating of Effective” would exemplify this approach. In this case, the underlying belief is that all hurdles are equally important and must be surmounted to reach a specified level of effectiveness (Mehrens, 1998). In contrast, when a compensatory approach is applied, higher performance on one component of the system (e.g., rating on student growth) is allowed

to compensate for lower performance on another (e.g., observation rating). The belief underlying this approach is that the different measures or components of the system are linearly related (Mehrens, 1998), such that a deficiency in one skill can justifiably be compensated by an excess in another.

The approaches outlined above are typically applied in practice using one of two methods for combining or aggregating system-level data: 1) computing a simple or weighted index, and 2) specifying a decision matrix. The index method orders the performance of teachers by aggregating scores within and between measures and/or components. States and districts using a composite or index score to determine final ratings for teachers include Colorado, Washington DC, New Jersey, Tennessee and Houston. Typically, point values are assigned to results from each measure and a sum or a weighted sum is calculated with the most weight going to those measures that: a) have the highest level of reliability in the system; or, b) have the most value to teachers.

On the other hand, decision matrix methods used in places such as Hawaii, Denver, Rhode Island, Wyoming, Georgia, New Hampshire, Wisconsin, and New Haven (CT), apply a rule-based approach to aggregate and assign teacher ratings at the component or system level. In this case, a chart or matrix is generated to illustrate how different combinations of scores or ratings provide for different results or system level categorizations. For example, the table below reflects the decision matrix used to award a final effectiveness rating to teachers in the state of Rhode Island (e.g., Highly effective, Effective, Developing, and Ineffective). The scores on the horizontal axis indicate a teacher’s rating on the student outcomes component of the system, while the scores on vertical axis reflect the rating for teaching practice.



		STUDENT LEARNING			
		4	3	2	1
PP x PF	4	HE	E	D	D*
	3	HE	E	D	D
	2	E	E	D	I
	1	D*	D	I	I

Figure 3. Example of a Decision Matrix developed by the Rhode Island Department of Education. Source: <http://www.ride.ri.gov/TeachersAdministrators/EducatorEvaluation/GuidebooksForms.aspx>

With respect to an overall rating of Effective, the table in Figure 3 clearly reflects a conjunctive approach; minimum scores of 2 and 3 (for teacher practices and student outcomes, respectively) must be obtained to be rated as Effective. The conjunctive approach is not, however, applied for the other classifications. For example, a high score in the area of student outcomes (i.e., 3 or 4) may compensate for a low score in the area of teacher practices (i.e., 1) to provide for a rating of Developing. In most states/districts, using a combination of methods and approaches such as that shown in Figure 3 is quite common.

Goals of Data Aggregation

When the purpose of evaluation is administrative, the primary goal of data aggregation is to establish an overall measure that is technically sound and provides for the classification or ranking of employees in a manner consistent with the decisions it is intended to support. It is important to note, however, that what constitutes a “technically sound” measure is a value-based decision, and will vary in light of the statistical characteristic desired from a final score or rating. For example, a composite measure reflecting overall educator effectiveness may be calculated by: 1) weighting measures or component scores in light of their estimated reliability (TNTP, 2010); 2) applying a set of “optimal” weights that serve to predict an identified key dimension of teaching or a target criterion (e.g., ability to influence student growth) (Mihaly et al., 2013; Glazer et al., 2011); 3) weighting measures in a way that provides for maximum differentiation or variability among educators; or 4) utilizing a decision matrix that ensures the overall score or rating is assigned in a manner that minimizes the likelihood of misclassification. In each of these cases what is valued from a statistical perspective differs and is reflected accordingly in the data aggregation procedures that are used.

At the developmental end of the evaluation spectrum, the primary focus is less on achieving high technical standards, and more on providing accurate and detailed information to employees regarding where they exist along a defined continuum of expectations specific to their job. Consequently, the goal of data aggregation is validation through triangulation, so that rich, contextually grounded information can be provided to employees regarding key areas of strength and need. Within this context, qualitative processes and tools are used (e.g., consensus moderation approaches to calibrate peer reviews) to evaluate a body of evidence, assess employee strengths and weaknesses and ultimately rate performance. Additionally, to sustain developmental goals, stakeholders play a significant role in informing how measures and components should be aggregated, reported and interpreted to support the provision of useful information to educators.

Unfortunately, in many systems PLDs that help communicate the expectations associated with different levels of performance - relative to a given measure, component, or the system overall - are either absent or poorly defined. This can result in the use of data aggregation procedures that are inconsistent with the goals of the system and, potentially, complicate the interpretation of system-based results. In the section that follows, we provide guidelines to support the development of appropriate, useful PLDs.

Considerations Related to the Development of PLDs for an EES

Regardless of the intended purpose of an EES, in order to establish meaningful performance standards, performance level descriptors (PLDs) that describe what educators must do to achieve a particular rating, performance level or classification category must be well articulated. This is true not only at the overall effectiveness level, but for each component or measure for which a performance standard must be defined. Specifically, PLDs must describe the intent of each threshold and the levels it defines in a way that will clearly be understood by all stakeholder groups. This is true for a variety of reasons. If performance expectations are not documented or clearly defined:

- The requirements necessary to achieve a given performance level or rating may be interpreted or operationalized differently across evaluators and educators, adding more subjectivity to an already highly judgmental and controversial process.

- Stakeholders will not know what knowledge/skills they are being evaluated against and therefore how to set appropriate professional goals.
- Stakeholders will not understand what it means to be rated within a particular level or category.
- The purpose or role of different components/measures within the EE system will not be transparent to stakeholders.

Furthermore, it is only by clearly defining what is expected at different levels of performance that supervisors will be able to make fair, consistent distinctions in performance when rating employees at the end of the appraisal period (“OPM.Gov”, n.d., Performance Management Cycle). Given the important role PLDs play in supporting the goals of educator evaluation, it is important they be of the highest quality, as defined by the following criteria:

- Fair – expectations should be reasonable, attainable and defined in terms of factors under the educator’s control
- Valid – expectations should reflect the knowledge, skills, competencies, or type of performance (e.g., growth, status, etc.) addressed by the measure or component to which they are associated
- Hierarchical – each performance level should reflect a logical, defensible progression of skills, attributes, or expectations
- Stakeholder-defined and approved – expectations should be viewed by stakeholders as reasonable and useful given the manner in which they are intended to be used (e.g., support standard setting, provide feedback to educators, inform higher-ed. as to expectations for incoming teachers)
- Coherent – descriptors provide for the type and quality of information necessary to support the goals(s) the measure/component was selected to inform
- Easy to understand– PLDs make use of language common to most educators, consistent with that reflected in state standards, or used in everyday conversation related to curriculum, assessment and instruction
- Quantifiable, Observable, and Verifiable² – expectations are defined in a manner that provides for evaluation and quantification of results in light of known resources

Meeting the criteria outlined above is not an easy task in any context; however specific challenges arise depending on the inferences and uses the standards must serve. Because they are 100% empirically derived, performance expectations defined to serve normative or predictive inferences are often easier to *write* than those developed to serve criterion referenced inferences. For example, within the context of growth, a “meets” performance standard could be defined as follows: “*An educator achieved a median growth*

² See: <http://www.opm.gov/policy-data-oversight/performance-management/performance-management-cycle/developing/differentiating-performance/>

percentile at or above that earned by 50% of the educators in the state the previous year.” In this case the fairness, validity, and coherence of the expectations are defined by the standard setting process (i.e., the data reviewed, stakeholders involved).³

In contrast, PLDs intended to serve criterion-referenced inferences, such as the development of personal improvement goals, require detailed explanation of the competencies, skills or attributes expected at each level. As a result, PLDs must be developed by stakeholders who clearly understand the construct underlying the standard, how it progresses, and the manner in which it is distributed in the population of interest to meet the criteria outlined above. In addition, PLDs should be crafted, vetted and finalized *prior* to engaging in the standard setting process, as previously discussed (Perie, 2008; Bejar et al., 2007; Egan et al., 2011; Ferrara et al., 2009).

Establishing Performance Standards for Administrative and Development Purposes

Once PLDs are articulated, appropriate standard setting procedures can be defined. This requires consideration of the purpose for evaluation, in conjunction with the PLDs and a variety of contextual factors that may influence the standard setting approach. For example, when administrative purposes are emphasized, performance standards should be defined in a way that promotes stakeholder confidence in the use of system-based results for driving human capital decisions. Consequently, standard setting decisions will be based largely on quantitative indicators. Standard setting panels may be asked to consider the percentage of educators expected to fall in each rating category and provide feedback related to the reasonableness of proposals yielding different distributions. Similarly, technical experts may be asked to evaluate sets of proposed standards in consideration of data reflecting reliability, accuracy, and decision consistency. To defend the standards as defensible and fair (for all employees) data presented should include that related to the expected impact of the proposed standards for different groups and/or information about the relationship between performance at the standard and one or more target criterion (i.e., other measures considered important to making decisions about effectiveness).

Standard setting activities to support developmental purposes should focus on defining performance levels and cuts that provide feedback to teachers regarding where their performance falls relative to expectations. Such information can facilitate meaningful discussions between educators and administrators and help target areas for improvement and professional development. Differential impact data will be evaluated, but their influence will be secondary to that of ensuring the standards align with the PLDs. Within this context, stakeholder contribution is necessary to ensure that performance expectations are clear, relevant, fair and reasonable; and that different levels of performance, when specified, are articulated in a manner that serves to both help educators understand where they fall relative to expectations and what they need to do to advance from one level to the next.

The Role of Context

In addition to the overarching purpose of evaluation, there are a variety of contextual factors that need to be considered when planning for and defining a standard setting process. Some of the most common factors are provided in Table 4. For each factor, the key question to be addressed and the potential impact of the response on the standard setting design is provided for consideration.

³ While an expectation such as that reflected in the MGP example clearly serves to explain the criterion for performance which must be met, to facilitate stakeholder acceptance, the rationale underlying this expectation and the process by it was defined, should also be described. This could be done briefly within the text of the PLD, or in a brief companion document (i.e., a subset of the Standard Setting Report.).

Table 4. Contextual Factors that Influence the Standard Setting Design

Factors	Key Question	How influence design of standard setting
Data combination method underlying the measure/ component upon which standards are to be set	What is the nature of the data upon which standards are to be set - scores, ratings, classifications or some combination of these?	Determines the nature of the standard setting activity (e.g., identifying threshold points on a score scale, defining rules that define the end result of a decision matrix) Determines the role of performance expectations and PLDs.
Inferences or Interpretations to be Made	What type of inferences do you want to make in light of observed performance relative to the standard? Criterion-referenced/Norm referenced/Predictive	Determines the data and materials necessary to support the setting of performance standards and the degree to which stakeholders are involved in the process.
Stakes associated with established standards	What are the stakes associated with the decisions that will be made in light of educator performance relative to the standards?	Influences the technical quality and the nature of the data provided for consideration, and who is involved in the standard setting process.
Timelines for implementation	When are the standards to be established? Prior to or after operational implementation?	Informs data that will be available for use to support the standard setting process
Data Availability	What data will be available to inform the standard setting process? To what extent is data privacy a concern?	Influences the type of data that will be available during and after standard setting to evaluate impact and validate appropriateness
Degree of Flexibility in Implementation	To what extent are the procedures associated with the measure/component of interest dictated across districts?	Influences the type and amount of data available to support the standard setting process (i.e., may differ from district to district). Influences the extent to which one can assume scores/ratings have similar meaning/impact/importance across districts. Informs the range of materials, necessary for review during the standard setting process.

Before defining any standard setting process, the nature of the data upon which the standards will be set must be clearly understood. For EES, this is determined in large part by the number of measures that must be considered and the process by which those measures are combined. In the simplest case, there is one measure (e.g., a teacher value added score) reported on a discrete or continuous scale upon which standards must be set. If multiple measures must be considered simultaneously, however, the method used to combine those

measures largely determines the type of standard setting approach that should be applied. For example, when an index or composite method is used performance standards must be set on the resulting, aggregate score scale. On the other hand, when measures are combined with a decision matrix, performance standards result from a rule-based standard setting approach. In this case, stakeholders review the ratings resulting from two or more discretely defined components/measures and generate a set of rules for combining the results to establish a final performance category, score, or rating for each teacher.

In addition to being influenced by the chosen method for combining data, standard setting approaches are also greatly influenced by the type of inference (e.g., norm-referenced, criterion-reference, predictive) desired. If, for example, the goal is to establish standards that allow for inferences about educator performance relative to the performance of others, data that accounts for the performance of this “norm” group will need to be factored into the standard setting process. Similarly, if performance standards are intended to support predicted inferences, such as the likelihood of a certain event occurring or criterion measure being attained, the standard setting methodology will need to supply the data necessary to support these inferences and provide direction to panelists as to how they should be interpreted and used. Such considerations will become more and more complex as the array of factors influencing the standard setting design increase.

To clarify how standard setting procedures are shaped by contextual factors, Attachment A describes what a standard setting might look like for a hypothetical EES similar in design to that reflected in Figure 1. We note that this scenario is not intended as an exemplar, and acknowledge that other procedures could be recommended. We provide this scenario primarily as a means of illustrating why one common standard setting approach cannot be prescribed for all EES.

PLDs and performance standards should support the purpose and goals of the EES, and be developed in consideration of intended or applied data aggregation techniques. If an EES does not establish a clear connection between these elements, the fairness and coherence of the system will suffer. In the final section of this paper, we describe how the process of developing PLDs and performance standards can contribute to the fairness and coherence of an EES.

Standard Setting to Facilitate Achievement of Coherence and Fairness in an EES

Coherence is noted by many educational researchers as necessary to help validate the design of an EES (Harris, 2013; Bell, 2012). Coherence dictates that the design of the system must clearly align with the state’s overall goals and purpose for evaluation, and that this alignment be both transparent and logical. In addition, the design of the system should clearly represent what the state values in terms of achieving the desired change. Procedures used to establish performance standards can contribute to efforts to achieve coherence in EES by ensuring that the data and materials used to establish performance standards are consistent with: 1) the goal or purpose the standard is intended to serve; 2) the type of inferences to be made; 3) the manner in which results are to be used, and; 4) the states overarching theory of action (i.e., hypothesis) as to how information resulting from the standard setting process will support the attainment of defined goals. When a standard setting process achieves coherence, the activities stakeholders engage in clearly reflect the manner in which the measure is *intended* to support the attainment system goals.

For example, assume there are two states (A& B), each with the same developmentally-based goal of improving classroom practices, but with differing theories of how the system will reach those goals. State A believes that the teaching practice measure will improve instruction by providing detailed information about an educator's strengths/weaknesses relative to state-defined expectations which can be used to target appropriate PD. On the other hand, State B believes that the teaching practice measure will support instruction by facilitating the development of professional learning communities that pair educators demonstrating strong teaching practices (i.e., those scoring in the top 25% within their school) with those who are struggling (i.e., those in the bottom 25%). Each state reports an average score resulting from the application of a 4-point teacher observation rubric on 3 separate occasions. However, the procedures used to establish performance standards would clearly differ for these 2 states. For State A, the goal is to make inferences about an educator's performance relative to clearly defined expectations and provide descriptive results and feedback about what he/she needs to do to improve. This necessitates the development of clearly defined PLDs, and a process that utilizes stakeholders to map those expectations on to the classroom practice scale. For State B, the goal is to establish a score and set of standards that allow for the rank ordering of educators based upon practices assessed so that they can be organized into effective teams. In this case, PLDs serve only to operationalize what it means, from a normative standpoint, to fall within a given level. The role of stakeholders is to identify the percentages that should be used to define strong and weak performance within a given school and determine how the norm group should be defined.

In practice, EES have multiple goals that are each informed by multiple measures/components. Therefore the picture quickly becomes complex. For example, some measures may be deemed more important than others at supporting a particular goal. As a result, defining an appropriate standard setting process for a given measure, component, or overall rating requires: identifying and prioritizing the goals a set of standards is intended to inform; outlining the type of information necessary to support those goals, focusing on the design features outlined in the EES; and then determining where inconsistencies or gaps exist. Clearly, not all goals will be equally informed by the same set of performance standards, but all goals should be informed by the information resulting from at least one set of performance standards.

Facilitating Fairness

The procedures used to support data aggregation and standard setting should establish performance standards that are appropriate for all teachers, regardless of the students they teach or the systems with which they are affiliated. Within the context of data aggregation, this means making sure that measures are combined in a manner that does not systematically disadvantage certain groups of teachers. For example, assume a state decided to calculate a composite using score reliability to determine differential weights across measures. If the reliability of the aggregate measures differs significantly for different groups of teachers (i.e., those associated with students with disabilities, working at low socioeconomic status (SES) schools, etc.) the resulting composite will be more error-laden for some teachers relative to others. Similarly, if the likelihood of missing data is higher for certain types of teachers data aggregation techniques that utilize those measures may be unfair to certain educators.

Within the context of standard setting, fairness involves consideration of whether expectations for performance are appropriately defined for all groups of educators in terms of both content (what is expected) and degree (how much). From a content perspective, it is

important that defined performance expectations be attainable by all educators (i.e., regardless of status, race, population taught, and school affiliation). That is, they should not reflect or be contingent on any factors that are out of an educator's control. For example, expectations that rely on educators having access to particular tools or resources may introduce a bias against those working at low resource schools. This is problematic not only from a fairness perspective, but also because it serves to disincentivize educators from working at the schools where they are needed most.

Similarly, expectations related to “how much” an educator must do to achieve a given performance standard must be fair for all educators. For example, if common performance standards are to be defined for educators that reflect an expected amount of growth (i.e., as reflected in growth based measures) all educators should have an equally likely opportunity of achieving those standards, regardless of who or where they teach (e.g., gifted/talented, SWD, etc.). This requires understanding the interaction between the measures used to inform decisions and different teacher/school-based characteristics, prior to the use of such standards. To be clear, this does not mean that if observation scores are found to be systematically lower at schools serving high poverty populations that these scores should be automatically invalidated. Rather, this finding should first trigger an investigation to triangulate the information against other sources and to identify plausible reasons for these results.

If a state or district is in a scenario where standards must be defined in advance of operational implementation, it is highly recommended that this be done with the understanding that validation activities using operational data will be required. As a means of meeting legislative requirements, or providing schools/districts with a clearer idea of what expectations for educator performance look like, many places have attempted to establish performance standards for their educator evaluation systems prior to implementation. Although often good intentioned, establishing performance standards in the absence of operational data can be extremely problematic because there is no way of knowing how system measures will perform (independently and in relationship to other measures) until full implementation. This is especially the case when dealing with a new evaluation system for which the reliability and validity of component measures is still in question, and for which procedures are still being piloted, evaluated and modified.

One of the many concerns associated with establishing performance standards in the absence of data (i.e., based only on written expectations for performance) is that results, when obtained, may not differentiate among educators in the manner, or to the degree, expected. This could occur in light of factors related to implementation of the evaluation process (e.g., evaluators are hesitant to give teachers low professional practice ratings, so all teachers obtain a score at the highest level) and/or unanticipated statistical factors related to the calculation, rounding and/or weighting of individual, or component measures (i.e., teacher practice, growth, or SLO-based measures). Consequently, if cut scores are established in the absence of performance data, there is a risk of severe misclassification of teachers.

The issue is similar to one of establishing cut scores for a test comprised of items and item formats that have not yet been field-tested. In this case, you know nothing about how the new items will perform (both in general and in specific sub-groups/populations) or how test scores might be distributed within your student population. If the items do not function as intended and/or the assessment fails to differentiate student ability, inferences based on the classification of students into performance categories may not be useful or valid. For example, if the items are unexpectedly hard due to issues related to quality, format, or opportunity to learn, all

students may be inappropriately classified in the lowest performance category – implying greater consistency in performance and lower proficiency than what truly exists within the population.

Although many states pilot components of their system in an attempt to gain an understanding of how system-based measures might perform prior to operational implementation (e.g., educator practice), a variety of factors may come into play during piloting that threaten the integrity and appropriateness of this data for use in defining performance standards, including factors related to: the efficacy of the evaluation process, the collection of required data (e.g., timing and attainment issues) and the availability of qualified resources. These may be general issues that affect all systems, or issues confined to particular schools and/or districts.

Furthermore, in many cases, pilot administrations are conducted using small, voluntary samples for the purpose of obtaining feedback regarding the clarity and quality of procedures and materials used. Preliminary data from these voluntary samples may be useful to support initial discussions related to the properties of measures and how they might be related, but will not be sufficient for setting high-stakes standards that impact educators – especially when there is a need to also evaluate impact by sub-group.

It is important to note that many states will be constrained in terms of what, and how much, EES data is made available to support standard setting and/or validation activities. This is due in large part to concerns around data privacy that have led many states to include language around access to educator evaluation data in law and regulation. To support accountability reporting, state departments may be provided with a list or summary of the final effectiveness ratings associated with a given school or district, but the component scores or measures that make up that evaluation are protected from release⁴. Clearly, this too will have implications for the manner in which performance standards are established and evaluated in an EES.

Conclusion

To ensure coherence and help address concerns about transparency and fairness, any efforts to differentiate performance across teachers using data aggregation methods should clearly articulate the expectations underlying the approach and the resulting performance standards. In places, such as New Jersey, where efforts are made to involve stakeholders in the standard setting process, performance standards and PLDs are shared across multiple entities and clearly connect back to the goals, purposes, and the theory of action set for the EES. But in the case of many other places where the standards were set in the absence of data, the onus is on the district or the state to make a compelling argument as to why the performance standards are reasonable and fair and how these standards contribute to the overarching goals of the EES.

As stated throughout this paper, the inclusion of PLDs and performance standards are critical elements to help achieve transparency in an EES and may help a state or district meet important criteria such as fairness and simplicity (Harris, 2013) in an EES. Within the world of performance management in the human resources sector, organizations such as SHRM and ANSI note that specific detail about behaviors, skills and measurements need to be defined to set clear performance standards or levels in a personnel evaluation system (SHRM & ANSI, 2012, pg. 24). Regardless of whether administrative or developmental purposes are emphasized in an EES, if

⁴ See: http://www.edweek.org/ew/articles/2012/03/28/26evaluation_ep.h31.html



performance standards are not supported by PLDs, policy makers will have difficulty communicating the quality and utility of system results to stakeholders.

Ultimately, states and districts bear the onus of having to demonstrate to stakeholders that their systems are meeting the stated purpose of an EES. Although we would prefer to see developmental goals emphasized in an EES, mainly because an administrative emphasis lacks a clear connection to a human capital and development systems theory, neither of these purposes can be simplistically and narrowly achieved based on formulaic or purely measurement decisions.



References

- American Standards Institute and Society for Human Resource Management. (2012). Performance Management, An American National Standard for Human Resource Management. Alexandria, VA: Author. Retrieved from: http://www.shrm.org/hrstandards/documents/12-0794%20performance%20mngmt%20standard_interior_viewonlyfnl_rvsd10-4-13.pdf
- Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 1–30). Maple Grove, MN: JAM Press.
- Bell, C.A. (2012). Validation of Professional Practice Components of Teacher Evaluation Systems. A paper presented at the 2012 Reidy Interactive Lecture Series, Boston, MA.
- Burling, K. (2012). Evaluating Teachers and Principals: Developing Fair, Valid, and Reliable Systems. Retrieved from: <http://educatoreffectiveness.pearsonassessments.com/>
- Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, 15, 12-21.
- 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. (2011). A validity framework for defining proficient performance and setting cut scores for accessible tests. In S. Elliott, R. Kettler, P. Beddow, & A. Kurz (Eds.), *Accessible Tests of Student Achievement: Issues, Innovations, and Applications*. New York: Springer.
- Ferrara, S., Dubravka, S., Skucha, S., & Murphy, A. (2009). *Test Development with Standard Setting in Mind*. Paper presented at the annual meeting of National Council on Measurement in Education (NCME) in San Diego, CA
- Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D.O., Whitehurst, G.J. (2011) *Passing muster: Evaluating teacher evaluation systems*. Washington, D.C.: Brookings Institute. Retrieved from: <http://www.brookings.edu/research/reports/2011/04/26-evaluating-teachers>
- Gordon, R., Kane, T.J., Staiger, D.O. (2006) *Identifying effective teachers using performance on the job*. Washington, D.C.: Brookings Institute. Retrieved from: http://www.brookings.edu/~media/research/files/papers/2006/4/education%20gordon/200604hamilton_1.pdf
- Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000a). Handbook for setting standards on performance assessments. Washington, DC: Council of Chief State School Officers.
- Hambleton, R. K., & Pitoniak, M. (2006). *Setting performance standards*. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.
- Hansen, M., Lemke, M., Sorensen, N., (2013) Combining multiple performance measures: do common approaches undermine districts' personnel systems? Washington, DC: American Institutes for Research. Retrieved from: http://www.air.org/files/VAMS/Combining_Multiple_Performance_Measures.pdf



- Harris, D. (2013) *How might we use multiple measures for teacher accountability?* Standard, CA: Carnegie Foundation for the Advancement of Teaching. Retrieved from: http://www.carnegieknowledge.org/briefs/multiple_measures/
- Kane, M. (in press). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Publishers.
- Lewis, D.M. & Green, R. (1997, June). *The validity of performance level descriptors*. Paper presented at the National Conference on Large Scale Assessment. Colorado Springs, Co.
- Lewis, D.M., Mitzel, H.C., Green, D.R. (1996). *Standard setting: A Bookmark approach*. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment: Phoenix, AZ.
- Mihaly, K., McCafferey, D., Staiger, D.O., Lockwood, J.R. (2013) *A Composite Estimator of Effective Teaching (MET Project Research Paper)*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from: http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf
- Mills, C.N., and Jaeger, R.M. (1998). Creating description of desired student achievement when setting performance standards. In L.N. Hansche (Ed.), *Meeting the requirements of Title 1: Handbook for the development of performance standards* (pp. 73–85). Washington, DC: US Department of Education.
- Montgomery County Public Schools. (2012). *Teacher-Level Professional Growth System Handbook*. Retrieved from: http://www.mceanea.org/pdf/Teacher_PGS%20Handbook_2012-13.pdf
- Office of Performance Management. (n.d.) *Select Performance Management Cycle website*. Retrieved from [:https://www.opm.gov/policy-data-oversight/performance-management/performance-management-cycle/planning/developing-performance-standards/](https://www.opm.gov/policy-data-oversight/performance-management/performance-management-cycle/planning/developing-performance-standards/)
- Perie, M. (2008), *A Guide to Understanding and Developing Performance-Level Descriptors*. *Educational Measurement: Issues and Practice*, 27: 15–29.
- Pulakos, E. D., (2004) *Performance Management: A roadmap for developing, implementing and evaluating performance management systems*. Retrieved from: https://www.pdri.com/images/uploads/Performance_Management.pdf
- Schneider, M. C., Huff, K. L., Egan, K. L., Tully, M., & Ferrara, S. (2010, May). *Aligning achievement level descriptors to mapped item demands to enhance valid interpretations of scale scores and inform item development*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.
- Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for achievement tests*. Stanford, CA: National Academy of Education.
- The New Teacher Project (2010) *Teacher evaluation 2.0*. Retrieved from: <http://tnpt.org/assets/documents/Teacher-Evaluation-Oct10F.pdf?files/Teacher-Evaluation-Oct10F.pdf>



Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009) The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. Brooklyn, NY: The New Teacher Project. Retrieved from: <http://widgeteffect.org/downloads/TheWidgetEffect.pdf>

Wingfield, E. (2013, November 13). Microsoft Abolishes Employee Evaluation System, *New York Times*. Retrieved November 15, 2013, from <http://www.nytimes.com>.



APPENDIX A:

Hypothetical Scenario: Standard Setting Considerations for State A

The scenario outlined below assumes an EES design similar to that described in Figure 2, and requires the specification of a variety of different types of performance standards to support aggregation, feedback, and ultimately the assignment of a final effectiveness rating. For each standard, a high-level process will be discussed that accounts for the overarching purpose of the system, data to be considered, inferences desired, theory of action, and the impact of specified contextual factors.

Purpose of the System:

The overarching purpose of State A's EES is to improve student learning by providing information and data to educators that informs instruction. While the overall effectiveness rating is also intended to *inform* administrative decisions related to promotion, retention and tenure, it is not on its own enough to result in the dismissal of an educator (i.e., additional evidence of the sort established through a committee review process is required).

Goals of the system:

The state's goals related to the development and implementation of the EES are predominantly developmental, and include the following:

- Improve teacher effectiveness
- Increase standardization across LEAs/schools with respect to how effectiveness is operationalized, discussed and evaluated
- Improve the quality and specificity of individualized professional improvement plans
- Improve teacher understanding and instruction of the Common Core State Standards (CCSS)
- Provide for gains in student learning
- Increase the number of new educators entering the field with skills necessary to be effective
- Establish stakeholder trust in the quality, fairness and relevance of the state's educator evaluation process.

Theory of Action:

The EES is intended to provide for the attainment of these goals by:

- Providing detailed and timely feedback to educators regarding performance
- Increasing opportunities for communication and collaboration between educators related to effective practices (in general and relative to the state standards);
- Increasing exposure to, and practice with, materials, procedures, strategies, and activities believed to be inherently beneficial to educators
- Providing for clarity and transparency in expectations for performance.

Summary of System Design:

A high-level summary of the design associated with the EES for State A is provided in Table A1. The system is comprised of four measures and two components which ultimately provide for an overall effectiveness rating. For each measure and component a process must be defined to establish the performance standards or decision matrices upon which final scores or ratings are based. Similarly, standard setting is necessary to determine how component ratings should be combined to obtain an overall effectiveness rating.

Contextual Factors of Relevance:

- Full implementation of the EES is planned for 2014-2015.

- The observation component of the system has been piloted, but not all components have been piloted in all districts – especially SLOs.
- Due to state laws related to data privacy, the SDOE will not be provided with teacher-level data. Instead, for each school, the state will receive summary-level data associated with each measure, component, and the final rating (i.e., the percentage of educators receiving each rating). The only exception to this is the value-added measure, which has been calculated by the state for each teacher for the past 5 years.
- With the exception of first year teachers (who require 3 formal observations) districts are able to determine their own rules related to the number of observations upon which educator practice ratings are based.
- Educators who perform at one of the bottom two effectiveness levels will be required to develop and adhere to a personalized improvement plan (PIP) and participate in 40 hours of PD.
- The state is transitioning their current assessment system to a system aligned to college and career ready standards. New assessments, fully aligned to these standards, will be administered for the first time in 2014-2015.

Table A1. Summary of EES Design for State A

Measures	Components	Final Rating
<p>Value-Added</p> <p>A value added measure (VAM) is calculated using a 3 year rolling average for all teachers who administer the state summative assessment.</p> <p>Final Growth Rating: Translate the value added score associated with an educator to one of four growth performance levels (e.g., Low, Moderate, Typical, High) using a table developed by the State Department of Education (SDE).*</p>	<p>Student Outcome Classification:</p> <p>For teachers in tested grades and subjects a 4x3 decision matrix (Growth Rating x SLO rating) is used to determine an educator’s overall</p>	<p>A 3X4 decision matrix considering Student Outcome Rating and Teacher Practice Rating and is used to assign Educators to one of four Effectiveness Levels (Not Effective, Partially Effective, Effectively, Significantly Effective)</p>
<p>Student Achievement on Teacher-Defined Learning Objectives (SLOs)</p> <p>Educators establish 2-4 student learning objectives (SLOs) and associated targets. For each SLO the degree to which a defined target is achieved is evaluated on a scale of 1-3, representing “Not Met”, “Met” or “Exceeded”</p> <p>Final SLO Rating: Combine scores obtained over all evaluated SLOs to establish a final SLO score.* Use this to place educators into one of three SLO performance levels (e.g. Unacceptable, Acceptable, Exceptional) given state defined score ranges.*</p>	<p>Student Outcome Rating on a scale of 1-3, representing “Low, Typical and High growth, respectively.</p> <p>For teachers in non-tested grades and subjects: Final SLO Rating = Student Outcome Rating.</p>	

<p>Student Perception Survey A state selected student perception survey is administered to all students at the end of the school year.</p> <p>Final Perception Rating: Calculate average score for all students associated with a given educator across grades and content areas. Assign an overall perception rating of 1-4*.</p>	<p>Overall Practice Rating:</p> <p>An educators overall practice rating is the weighted sum of their Perception, and Observation ratings, using the following equation:</p> <p>Overall $PR = .15(\text{Perception}) + .85(\text{Observation})$</p> <p>This results in a score in the range of 1-4, which is rounded to the closest integer for a final rating of 1-4.</p>	
<p>Observations: All educators in the state are observed in the classroom and scored using a common set of rubrics aligned to a state-specified framework for teaching.</p> <p>Final Observation Rating: Mean performance across occasions is calculated and state-defined score ranges are used to assign an overall Observation rating of “Unsatisfactory, Approaching, Proficient, or Distinguished”*.</p>		

Note: * = to be defined through standard setting

Value Added Measure:

The value-added measure (VAM) is intended to increase educator effectiveness by providing teachers with information about the impact they are having on student learning, as reflected by student performance on the state summative assessment. The performance standards are intended to support criterion-referenced inferences regarding the extent to which educators have achieved a level of growth consistent with that expected – as reflected by a rating of Typical or High.

Given the technical nature of the measure, State A plans to consult with their Technical Advisory Committee to establish an initial set of proposed performance standards for the VAM. This activity is intended to occur in summer of 2014, prior to full implementation. Subsequently, the proposed standards and a summary of the rationale by which they were established will be brought to a stakeholder panel for review and approval.

Technical advisors will be asked to evaluate historical value-added data, and what they know about the characteristics of the state VAM, to make an initial determination of how “typical” performance, as well as performance at the other levels, should be operationalized. To support this discussion, TAC members will be provided: an overview of the state value-added measure, the distribution of teacher value-added measures in the state (overall and by teachers associated with particular student sub-groups) for the last 3 years, and the reliability of the VAM estimate. Reliability will be considered both in terms of the degree of error within the estimate, and the consistency of ratings assigned to educators over years. Evaluating this information is helpful for determining how many performance levels can be reasonably supported. For example, if a large degree of error is detected, fewer performance categories should be set to avoid the possibility that teachers are assigned to categories based largely on chance. Once a set of proposed standards are identified the expected impact associated with those standards (based on historical data) and a summary of the rationale for their proposed placement will be presented to



a panel of stakeholders for review. The stakeholders will be provided with the option of shifting the proposed cuts within a range defined by the state if they feel it is necessary (e.g., to be fair or to better reflect educator expectations).

As the impact of transition to a new assessment on the state VAM is not yet known, the TAC recommended that any performance standards be considered preliminary until they can be evaluated using operational data.

Student Achievement on Teacher-Defined Learning Objectives (SLOs)

While the process of defining and developing SLOs is intended to support state goals related to improving effectiveness through increased communication and exposure to beneficial practices (e.g., setting and monitoring targets for student performance); SLO *ratings* are intended to support the goal of increasing educator effectiveness by providing teachers with information about their ability to define and achieve student learning targets within their classroom. In State A, educators submit 2-4 SLOs, each of which are independently scored, so a process must be defined to establish an overall SLO score (e.g., use of the sum or average) and assign it to one of three SLO performance levels.

SLO performance standards are intended to support criterion referenced inferences regarding the degree to which teacher-specified SLO-targets were attained. To set standards, State A's DOE has decided to convene a panel of stakeholders consisting of educators from a variety of disciplines. During the meeting educators will first be asked to articulate what SLO performance that has "Met" expectations should look like when evaluated across SLOs. That is, what factors, specifically, are relevant to consider? Should the state take the average score earned across SLOs or use a rule-based, decision matrix approach? Should consistency in performance across SLO targets play a role in the overall rating or should the best performance achieved across all the SLOs be applied? Panelists will be asked to discuss these different factors, thinking specifically about how they influence the manner in which expectations are defined, and come to consensus regarding how the SLO scores should be combined. Draft PLDs that reflect the chosen approach, will be drafted and refined throughout the discussion with the understanding that any rules outlined within will be provided to educators to support the interpretation of results.

Since the SLO process will only been piloted within a few districts across the state, summary data reflecting the impact of the proposed cuts will not be available for review as part of the standard setting process. (Even after 2014-2015, due to data privacy laws this information will not be provided to the state for review). Consequently, the focus of standard setting for this measure will be less on data, and more on identifying the type of performance (across SLOs) necessary to meet expectations as previously defined.

To inform the process, educators will be provided with a series of tables reflecting profiles of performance across SLOs and the different results that would accompany each of several options for combining the data. Educators will use this information in conjunction with the defined performance expectations to determine where the standards should be placed, and whether one set of standards is sufficient regardless of the number of SLOs associated with the rating.

Student Perception Survey:

Performance on the student perception survey is intended to provide feedback to educators on how students perceive their performance in the classroom. While a detailed report summarizing performance in domains related to instruction, temperament and classroom atmosphere will be provide to educators for review, the final perception rating is intended to signal whether the majority of students assigned to a teacher believe that he/she provides the type and level of support necessary for academic success.

To establish performance standards for the student perception survey, panels of educators will be convened to review performance profiles for teachers with low, medium and high perception scores. Profiles will be based upon teacher performance in 2 domains from the state



applied teacher observation framework, selected by instructional specialists who indicated that both domains focus on the quality of teacher engagement with students. Although restricted to the pilot districts, this preliminary standard setting exercise will help the state determine whether profiles of teacher practices can be used to inform the specification of cut scores for the student perception survey. The state intends to revisit the initial cut scores set once the full set of data is available.

Observations

To support the EES as intended, teacher observation ratings should provide educators with useful information regarding the manner and degree to which their practices reflect those expected by the state. In addition, they should clarify how effective teaching is defined in the state, so that current and incoming educators, parents, principals and state superintendents have a clear, common understanding of expected performance.

Given the way in which observation rubrics have been defined, including articulated performance expectations for each domain assessed within the framework (e.g., classroom instruction, classroom environment, etc.), State A has decided that the overall observation rating for a teacher will simply be the average observation rating earned within a given year. Since the number of observations associated with a teacher will vary depending on rules defined by the district, for some educators this value will not be an integer.

In light of the factors outlined above, the process of establishing performance standards for this measure involves two phases. In Phase 1, a panel of educators that span grades and disciplines will convene to generate comprehensive performance expectations that go across the domains targeted with the applied framework for teaching. To support this activity, panelists will be provided with the observation rubrics used to score performance in each domain, and the scoring rules used to aggregate performance across domains for each observation.

In phase 2, the panels will identify the percentage of educators that they believe should fall within each of the given performance levels given their beliefs as to how the skills/competencies reflected in the expectations are distributed in the state, and what would be considered fair and reasonable given the way in which results are to be used.

While most districts will have piloted the observation component of the system prior to full implementation in 2014-2015, only a small, voluntary sample of districts agreed to submit this data to the state for review and evaluation. This data will be used to identify values on the observation rating scale that provide for percentages consistent with panelist expectations, but the final cut score will not be determined until after operational implementation.

The Overall Student Outcome Classification

The Overall Student Outcome Classification is intended to reflect a teacher's ability to facilitate student growth and learning relative to that expected by the state or reflected in teacher-defined targets. Since the state believes that the growth reflected by and educator's SLO rating is *as important* as that reflected by VAM given the goals of the system, a 4x3 decision matrix (VAM x SLO) which considers the joint influence of both of the measures will be established to determine the Overall Student Outcome Classification.

The state's initial recommendation, which reflects a compensatory approach, is presented in Figure 5. Using this model, a high VAM rating is allowed to compensate for a low SLO rating, and vice versa.

Figure 5. Decision Matrix Associated with State A Student Outcome Classification

VAM	SLO Rating		
	Unacceptable	Acceptable	Exceptional
Low	L	L	T
Moderate	L	T	T
Typical	T	T	H
High	T	H	H

L=Low; T=Typical, H=High

The matrix was defined by the state in consideration of historical VAM data in conjunction with expectations about the percentage of educators that might obtain each SLO rating given: discussions with stakeholders, research related to SLOs, and data observed in other states.

To build shared ownership of this system with stakeholders, the state will convene committees of educators to comment on the proposed matrix. Stakeholders will be asked to first review the expectations associated with the SLO and VAM rating and then define what “Typical” growth should look like within the context of these two measures. Subsequently, stakeholders will be presented with the matrix, the rationale behind its development and the data considered during specification. Based on these discussions modifications may be recommended.

Overall Practice Rating

The Overall Practice Rating (1-4) is intended simply as a summary rating that accounts for both teacher observations and the result of the Student Perception Survey. The Outcome rating is calculated as a weighted sum of these two measures that is subsequently rounded to the closest integer value. For this component, the standard setting process is reflected mainly by the data and materials considered by the state to establish the weights applied to each measure and how/if the resulting values should be rounded.

To support these decisions, stakeholders were asked to provide feedback as to the degree to which each measure should influence the overall practice rating, given the type of information reflected by each and the state’s overarching goals of improving teacher effectiveness and increasing student growth. Subsequently, the State DOE worked with its Technical Advisory Committee to consider these recommendations, to determine how different weights and rounding rules might influence the percentage of educators receiving each rating. Through this process, it was determined that the final perception ratings should represent 15% of the overall practice score and the resulting index value should be rounded to the closest integer to determine the overall practice rating.

Next Steps

The final step in this hypothetical scenario would be to continue the efforts of building shared ownership in this system by having stakeholders move through a similar standard setting process for the final teacher rating, and validating *all* proposed standards once operational data are available. In the scenario outlined here, the procedures used to recommend performance standards are limited by contextual factors, but still meet the criteria defined in this paper as necessary to establish a fair, transparent and defensible EES.