A Framework to Support the Validation of Educator Evaluation Systems

Erika Hall, Ph.D.

National Center for the Improvement of Educational Assessment

Introduction

In September 2011, the U.S Department of Education (USDE) provided states with the option to submit a waiver to obtain flexibility related to certain provisions of No Child Left Behind (NCLB). In order to receive this flexibility, however, states were required to describe their plans for committing to the following four principles:

1. Adopt college- and career-ready standards in Mathematics and English Language Arts (ELA) and evaluate student performance relative to these expectations in (at least) grades 3-8 and once in high school.
2. Develop and implement differentiated systems of accountability with support focused on: improving academic achievement for all students, closing achievement gaps, and improving equity.
3. Promote teacher quality and principal leadership by implementing evaluation systems that utilize multiple measures of performance (including student achievement) and provide feedback that supports professional development initiatives.
4. Remove duplicative and burdensome reporting requirements that have little or no impact on student outcomes.

Overwhelmingly, states applied for and were granted NCLB waivers. As of February 2014, 42 states as well as the District of Columbia and Puerto Rico received approval of their flexibility requests[1].

Although most states had teacher evaluation systems in place prior to applying for flexibility; many were required to make significant modifications to meet the specifications in principle #3. These specifications include utilizing multiple measures of performance, including professional

[1] See the full ESEA Flexibility Policy document at the link: http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

practice measures and those that reflect student growth; providing teachers with feedback that informs professional development; and producing information that supports administrators in making personnel decisions[2]. In most cases, compliance has proven to be a long, arduous process requiring ongoing input from stakeholders and technical advisors, and constant evaluation of the appropriateness, fairness and relevance of proposed procedures, measures and results.

Due to the complexity of many new educator evaluation systems (EES) and the novelty of the techniques and measures used to support their implementation, several states opted to develop and pilot the components of their system in a step-wise fashion with the goal of incremental improvement prior to full implementation. Although this helps states and districts understand the strengths and short-comings of individual system components and identify where modifications may be necessary, such practices are problematic because they delay analysis of the system as a whole until after it goes live.

Although the USED recently relaxed the start date for states to begin using student growth in support of personnel decisions until 2016-2017[3], for many states this deadline remains a huge hurdle as the infrastructure, support and guidance necessary to assist in the development and implementation of these systems far surpasses the resources many state departments of education (SDEs) have in place. Even for states with fully operational EES prior to 2016-2017, there will often be measures that require multiple years worth of data before being considered reliable enough to contribute to an educator's overall effectiveness rating and/or be used for accountability purposes (e.g., using a 3-year rolling average for value-added measures). In addition, given the need to adopt new college- and career-ready standards, many states are still in the process of assessment and accountability transition. For these systems, the relationship between and among components that rely on student-outcome data will remain partially hypothetical until multiple years of implementation have passed.

[2] See: Principal #3, on pages 2-3 of the ESEA Flexibility Policy document, at the link in footnote #1

[3] See additional request for flexibility letter from the Assistant Secretary at: http://www2.ed.gov/policy/elsec/guid/secletter/080713.html

Due to these and other challenges it is clear that many states' teacher evaluation systems are being installed without adequate validity studies beforehand (Sheppard, 2012). They are based in large part on *theories* as to how the system will bring about change, what motivates teachers, the extent to which measures provide for reliable, relevant information related to the constructs of interest, and the efficacy with which system-based procedures are being implemented. While validation is an ongoing process, such that the full array of evidence necessary to support these theories will not be in place prior to implementation, forging down this road without sufficient data can have long-lasting negative consequences. Should preliminary results run contrary to expectations, states may face a loss of stakeholder support and confidence in their system.

For example, in many states, part of the push to move to new EES was to counter criticisms that old evaluation systems were not effective at differentiating among educators, despite the commonly held belief that there is a great deal of variation in the quality of teaching represented in our schools (Weisberg, Sexton, Mulher, & Keeling, 2009; Burling, 2012). If stakeholders expect the new systems to detect such variability and have been told that this is the mechanism by which improvement will occur, a system which initially results in virtually all educators being rated in the same way may be condemned as useless and a waste of money. Unfortunately, this will be true regardless of the reason for the result (e.g., educators still learning how to implement components of the system) and despite any positive benefits the system may afford (e.g., increased communication among teachers).

More importantly, in most states EES are being used to make personnel decisions about educators in addition to collecting information to support improvement. If systems are put in place without sufficient evidence that these measures and the manner in which they are being aggregated provides for fair, reliable inferences, inappropriate decisions about the effectiveness of individuals or groups of educators may be made. Similarly, if preliminary results suggest that certain types of educators may be at a disadvantage, such as those working with low-ability students or students with disabilities (SWDs), the system could have the negative effect of discouraging high quality educators from working in disadvantaged schools or with those populations for which their support is most greatly needed (e.g., Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, & Shepard, 2010).

To help mitigate issues such as those above, many states have decided to hold off using results from their EES for accountability purposes so that they can review the characteristics of system-based measures and collect feedback from stakeholders regarding the quality and clarity of implementation. Although preliminary analyses such as these are important, to truly defend the use of system-based results, comprehensive validity studies based on the collection of evidence aligned to system-based claims must be defined and conducted for EES prior to implementation, once they are in place and for the years which follow.

Purpose

The intent of this paper is to present a framework that supports the development of a comprehensive design argument for EES using the principles of evidence-centered design (ECD) (Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003). This work builds upon the work of others by providing a structure that necessitates articulation of the claims, inferences, and assumptions underlying a given EES in such a way that the evidence and research necessary to support validation is clear and transparent.

To better illustrate why a framework is necessary, the section that follows describes a common design for EES and illustrates the range of ways in which two systems may vary despite perceived similarities in structure and format. It is followed by presentation of the framework, a definition of its elements and a set of guiding questions. The paper concludes by illustrating how the framework provides a clear, coherent foundation by which to frame a comprehensive interpretive argument and identify the range of evidence and analyses necessary to support validation.

Defining the Validation Effort

While most states plan to collect evidence to support the use of their EES, at this point very few are doing so in a cohesive, deliberate manner, such that the network of inferences, and assumptions necessary to support those inferences, are clear. Research questions are often defined in terms of what others have done, what seems "right" or "necessary", and/or data that is, or will be, readily available for use. While these are all important considerations, especially the

latter, a research agenda designed in this manner will not provide for the complex array of evidence necessary to support a specific system design.

Within the context of educational assessment, validation is considered a process of compiling evidence to support one's argument that the information resulting from an assessment provides for fair and accurate inferences aligned with the defined goals and purpose for testing (Kane, 2006; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Although EES are not assessments, per se, most are comprised of components that can be considered unique (but not independent) "assessments" or measures of the constructs deemed necessary to support inferences related to educator performance. It is not surprising, therefore, that many experts (Bell, C., 2012; Sheppard, L., 2012; Bell, Gitomer, McCaffrey, Hamre, Pianta, & Qi, 2012) have recommended developing validity arguments for EES using models similar to those proposed for assessment systems.

Bell (2012) illustrated how an interpretive argument based upon Kane's (2006) approach could be used to validate the use of professional practice measures as a means of improving teacher instruction through feedback. In order to use the results as intended, such an approach requires articulation of the array of inferences and assumptions that must be met, followed by specification of the type of evidence that could be collected to support those claims. Similarly, Sheppard (2012) presented a theory of action (TOA) approach to outline a validity argument for the use of tests-based measures as a component in EES. Specifically, she summarizes the assumptions underlying the TOA for the summative and formative use of test-based measures of student growth, the rationale for those assumptions, and the research that has been conducted to support them.

The Design of Educator Evaluation Systems

Figure 1 presents a typical design for the educator evaluation systems currently specified across many states and districts to comply with ESEA waiver or Race to the Top (RTT) requirements. From bottom to top the figure represents the collection of Measures (purple) which are rolled up into Components (green), and then further aggregated to establish an Overall Rating (blue).

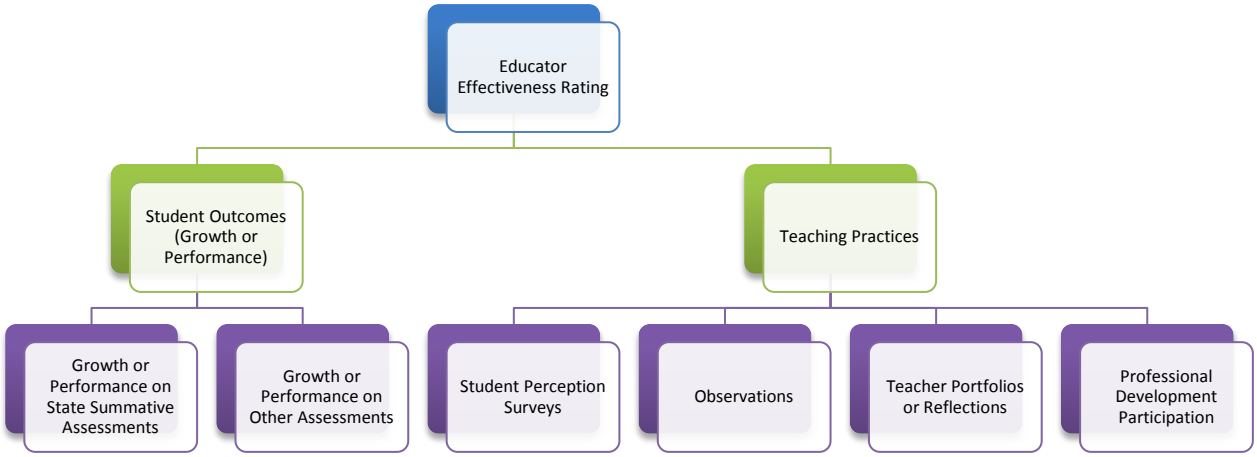

Figure 1. Common Design for an EES. From: Considerations for Establishing Performance Standards for Educator Evaluation Systems by E. Diaz-Bilello; E. Hall & S. Marion, 2014. A paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Within this model, the educator effectiveness rating is defined in terms of two components: Student Outcomes and Teaching Practices. This dual conceptualization represents the common belief that evidence of both performance-based factors (including expected behaviors and characteristics) and outcome-based factors are necessary to make fair, accurate and meaningful inferences about the performance of teachers within and across years. Each component is operationalized in terms of one or more typical "categories" of measures which are selected and defined by the state. For example, due to requirements outlined by USDE, all states/districts that received a flexibility waiver must include at least one student outcome measure that reflects the influence of an educator on student growth. Similarly, most systems necessitate the observation and evaluation of educators in the classroom as a means of collecting evidence of educator practice relative to defined expectations for performance. We understand that many variations on this design are possible, not only in terms of the components and measures specified, but also the manner in which they roll-up to support the specification of system-level

scores and ratings. For ease of discussion, however, this particular design will be referred to throughout this document.

At first glance, many EES are judged to be similar because they consist of components and measures that share the same name or label, such as those outlined in Figure 1. However, in most cases these systems differ greatly due to the way in which such elements are defined, prioritized, and measured. Even when components or measures are operationalized in a similar manner (e.g., Value Added Model (VAM) calculations are similar; use of same student perception survey, etc.) , states typically have different reasons for including them in the system that stem from different beliefs about the role they should play in bringing about desired outcomes. Consequently, when it comes to validation, using the work of others as a guide must be done with caution as it could lead to the collection of data that is irrelevant or even contradictory in light of one's defined goals or purpose for evaluation.

For illustrative purposes, consider two hypothetical states A & B. Each state has developed an educator evaluation system in which an educator's final effective rating is based upon a component related to student outcomes and a component related to teaching practices. Within each state, student outcomes are operationalized in terms of a measure of student growth based on the state administered assessment, and an additional measure that reflects student performance on key standards. Similarly, in both states teaching practices are operationalized in terms of an observation measure, a student perception survey and a compilation of evidence reflecting educator perceptions and practices in and out of the classroom. Table 1 summarizes, at a high level, how each measure is operationalized and represented to support reporting and aggregation for State A and State B, respectively.

Table 1

Description of Measures Included in Teacher Evaluation System for States A & B

Measure	*State A*	*State B*
Growth on State Summative Assessment	For each teacher a value added measure (VAM) is calculated using a 3-year rolling average.	For each teacher a median Student Growth Percentile (SGP) is calculated using all students associated with a given teacher in a given year.
	Applies to all teachers in grades 3-8 and 11 who teach Math or ELA and administer the state summative assessments.	Applies to teachers of 4-8 Math and ELA, Algebra I, Geometry and end-of-course Language Arts who administer the state summative assessments.
	Educators in non-tested grades and content areas do not receive a measure for this element of the system.	Educators in non-tested grades and content areas receive the median SGP associated with educators who teach tested grades/courses in their school.
	Final VAM Rating: Use state-defined score ranges to translate the VAM associated with an educator to one of four growth performance levels (e.g., Low, Moderate, Typical, High)	*Final SGP Rating***:** Use state-defined score ranges to translate the median SGP associated with an educator to a rating of 0-3 using state defined ranges.
Performance on Other Assessments	Educators in all content areas develop 2-4 student learning objectives (SLOs) and associated targets as a means of evaluating and measuring class progress relative to teacher-defined expectations for performance.	Educators in all content areas administer and score state-developed performance based tasks 3 times a year.
	For educators of ELA and Mathematics SLOs must be aligned to those Common Core State standards identified by State A as "focal" within a given grade and content area.	For educators of ELA and Math the PBTs are aligned to those Common Core state standards identified by State B as difficult to address effectively within the context of the state assessment
	For each SLO, educators receive a rating of 0-3 based on the degree to which the defined target is attained. Individual district administrators and teachers agree on rules for determining whether the target is: not attained at all (0),	For each task, educators receive a rating of 0-3 based on the percentage of their students who obtained a score of 2 or better. The state determines the percentage ranges associated with each rating: 0-15% (0), 16-40% (1), 41-60% (2), 61% or above (3)

Table 1 (cont.)

Measure	*State A*	*State B*
	partially attained (1), attained (2), or exceeded (3).) *Final SLO Rating*: Calculate the average score obtained over all evaluated SLOs. Use this to place educators into one of three SLO performance levels (e.g. Did not Meet, Met, Exceeded) based on state-defined ranges.	*Final PB Task Score*: Calculate the average teacher rating over all 3 performance task events. This results in a non-integer score between 0-3.
Student Outcome Component Rating	*For teachers in tested grades and subjects*: A 4x3 decision matrix (Growth Rating x SLO rating) developed through a state-defined standard setting process is used to determine an educator's overall Student Outcome Rating on a scale of 1-3. The matrix allows high growth to compensate for Low performance on SLO targets. *For teachers in non-tested grades and subjects*: Final SLO Rating = Student Outcome Rating.	*For all teachers*: Student Outcome Rating is calculated as the weighted sum of the Growth rating and the Performance-Based Task Rating using the following equation: Student Outcome=.75(growth)+.25(PBT) This results in a non-integer value in the range of 0-3 that is translated to a overall Student Outcome Rating of 1-4 using state defined ranges.
Student Perception Surveys	Teachers administer a state selected student perception survey at the end-of-the school year. *Final Perception Score/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 based on state defined score ranges.	Teachers administer a district developed or selected student perception surveys 2 times per year (mid-year and end of year). *Final Perception Score/Rating:* Calculate the change between an educator's mid-year score and end-of-year score. Assign an overall perception rating of 1-3 based on the degree of change observed relative to that expected given state defined norms.
Observations	All educators in the state are observed in the classroom on 3 separate occasions and scored using a common rubric aligned to a *state-specified* framework for teaching.	All educators are observed in the classroom using one of 4 state-approved, *district selected*, observation frameworks. The number of formal observations required is defined by the district.

Table 1 (cont.)

Measure	*State A*	*State B*
Observations	*Final Observation Score/Rating:* Mean performance across occasions is calculated and state-defined sore ranges are used to assign an overall Observation Rating on a scale of 1-4.	*Final Observation Score/Rating:* Districts assign educators to one of four performance levels based on their evaluation of the full range of information collected across all observations relative to state defined practice-based performance level descriptors. (e.g., Exemplary, Proficient, Emerging, Satisfactory)
Teacher Portfolios or Reflections	All educators are required to submit a portfolio of evidence they believe represents the nature and quality of their practices (e.g., beyond classroom instruction) over the course of the academic year. *Final Portfolio Rating:* A state-defined rubric is used to rate educators on a scale of 1-4 on the degree to which they demonstrate school responsibility, communication, professionalism, and planning as reflected through observation as well as the collection and review of submitted artifacts (e.g., lesson plans, syllabi, etc…) .	All educators are required to evaluate their performance over the given school year relative to expectations defined and approved at the beginning of the year in conjunction with their administrator. Educators must compile evidence to support claims related to the quality and completeness of defined goals. *Final Portfolio Score/Rating:* Administrators assign educators a rating of 1-4 based on their review of the compilation of evidence provided by the teacher, the teacher's Student Perception Rating, and a face-to-face discussion. Educators are rated on the extent to which provided evidence reflects the attainment of goals; Insufficient Evidence (1); Some Evidence (2); Sufficient Evidence (3); Substantial Evidence (4)
Teacher Practice Component Rating	An educators Overall Practice Rating is the weighted sum of their Perception, Observation and Portfolio-based ratings *rounded to the closest integer value*, as reflected in the following equation: Round(Overall Practice Rating) =.10(Perception)+ .55(Observation)+.35(Portfolio)	An educator's Overall Practice Rating is determined using a state-defined 4X4 decision matrix that considers an educator's Observation Rating and Portfolio Rating. Educators are assigned an overall practice rating of 1-4 to reflect unsatisfactory, partially satisfactory, satisfactory or distinguished performance, as defined by stakeholder developed PLDs.

Table 1 (cont.)

Measure	*State A*	*State B*
Final Effectiveness Rating	A Final Effectiveness Rating is assigned using a state-developed, stakeholder approved, 4x3 decision matrix of Teacher Practice Rating and Student Outcome Rating: Not effective, (1); Partially Effective (2); Effective, (3); Significantly Effective (4).	A Final Effectiveness Rating is assigned by calculating the average of the Student Outcome Rating and the Overall Teacher Practice Rating and then placing educators in one of four categories based on state-defined score ranges: 0-1.5 =Not effective; 1.5-2.25 = Partially Effective; 2.25-3.5 = Effective, 3.5-4 = Significantly Effective.

From this example it is clear that, although they share the same two components, the EES defined for State A and State B differ in a variety of significant ways, including

- the way in which the measures are operationalized and calculated (e.g., use of value added vs. student growth percentiles to support growth inferences);
- the manner and precision in which the results associated with a measure are represented to support reporting and aggregation (e.g., non-integer score, rating, performance level);
- the rules applied to teachers associated with tested vs. non-tested grades and courses;
- the process used to combine measures for the purpose of establishing a component or final effectiveness rating (weighted composite vs. use of a decision matrix; compensatory vs. conjunctive approaches);
- the weight given to different types of measures when defining a component or final effectiveness rating; and
- the flexibility afforded to districts in terms of how different components of the system are selected and implemented.

There are a variety of other factors not represented in Table 1 that would also reveal differences between these two models, such as: the process used to establish performance standards for a measure, component or system level construct; the extent to which stakeholders were included in the design specification process, and the tools, programs, supports, put in place to support implementation.

Given this, one may ask, "Why do they vary?" If two states have the same goal – to evaluate educators – and are using essentially the same components and categories of measures, why aren't their systems defined in a similar way? The answer to this question may, in large part, seem obvious: Different states have different tests, teaching standards, resources, school calendars, and legislative mandates that, in large part, drive how their systems are defined.

However, while these factors play a role, they probably do not result in as much variability as one may think. In addition, they do not explain why districts within a state that share these features often decide (if given the opportunity) to define their own EES – despite the significant cost, time and effort.

Ultimately, it is a state/district's unique values, priorities and beliefs related to the overarching purpose and goals of educator evaluation, in conjunction with policy and resource constraints, which has the *greatest* impact on decisions related to system design and implementation. Such variability is often reflected in how states/districts respond to key questions such as:

1. What is the **primary** purpose of educator evaluation (e.g., support administrative decisions support the development of educators, increase credibility)?
2. What are the short and long term **goals** associated with implementation of an EES and how are they prioritized (e.g., improve teacher effectiveness, show gains in student learning, remove ineffective teachers, increase the number of new teachers entering the field each year that meet effectiveness expectations, etc.)?
3. What are the **key processes/mechanisms** that will support the attainment of defined goals? Or, what are the **design features** that are most likely to bring about desired change? (Increased communication/collaboration among educators, increased motivation, focus on standards, provision of data, etc...)?
4. How is the **domain** of interest defined? What **constructs** are necessary to support inferences and decisions related to that domain?
5. What **data or measures** should be collected to operationalize, or quantify, the constructs of interest?
6. What is the **expected relationship** among the constructs of interest and the measures selected to represent them?

These questions indicate not only the state/district's impetus for evaluation, but also the overarching Theory of Action (TOA) as to how the system will support the attainment of system goals and the way in which the domain of effectiveness is defined. As a result, how a state responds to these questions determines how measures are selected and weighted, the types of procedures used to support data aggregation, and the inputs put in place by the state/district to support the attainment of system goals.

For example, if a state sees the primary purpose of their evaluation system as supporting high stakes administrative decisions related to promotion, retention, removal, etc…the procedures identified to support data aggregation may be defined in large part by the psychometric characteristics of system-based measures. Greater weight may be given to those measures having higher reliability, or those shown or expected to correlate more highly with a target criterion of interest (e.g., Mihaly, McCaffrey, Staiger, Lockwood, 2013; Hansen, Lemke, Sorensen, 2013; Glazerman, Goldhaber, Loeb, Raudenbush, & Staiger, 2011). Such practices would be consistent with that used by State B in calculating a Final Student Outcome Rating if, for example, performance based tasks were scored by teachers and were significantly less reliable than SGP measures for making inferences about an educator's impact on student outcomes.

In contrast, if the primary purpose of evaluation is to provide feedback to educators that can support instructional improvement, data aggregation may focus less on psychometrics and more on the triangulation of evidence across measures in a manner that facilitates accuracy in the identification of areas of strength and need (Dibello, E., et al., 2014). For example, in State A the Final Student Outcome Rating is assigned using a decision matrix defined by educators through a formal standard setting process. Such a process necessitates clarity in the expectations associated with performance at each measure and agreement as to the relative value and weight of each in making a statement about an educators overall ability to impact student growth.

Similarly, if a state believes that communication is the key mechanism by which an EES will provide for the attainment of system goals, they may select measures that necessitate teacher collaboration (e.g., SLOs) and/or provide for inputs that facilitate interdisciplinary discussion. In contrast, if a state's TOA relies largely on the provision of data-based feedback to inform instruction, the development of data maintenance and reporting systems may take priority to support educators in the timely access and interpretation of student achievement results.

While questions 1-3 require specifications of the overarching goals and intended uses of the system, questions 4-6 ask how the domain of evaluation should be defined and operationalized in order to achieve those goals. In her paper, Bell (2012) makes the distinction between teacher quality and teaching quality to illustrate how establishing clarity around the domain of interest and the constructs selected to represent that domain are necessary to support validation:

> Most states and districts want to make claims about teacher quality. But in general, we do not have measures of teacher quality. We have measures of teaching quality. The distinction between teacher and teaching quality concerns the degree to which we are measuring traits of teachers (teacher quality) or we are measuring traits of teachers, students and their contexts (teaching quality). The difference between the two originates in one's conception of the phenomenon of teaching. (p. 4-5).

She explains that due to the interactional nature of teaching and learning, most measures of student learning account for a variety of factors (e.g., characteristics of the teacher, the students, and the context in which they interact) that may support inferences related to *teaching* quality, but may not be appropriate to support inferences about the quality of a teacher independent of these factors— that is, across different groups of students, within different schools, or in light of differing resources and levels of support. To make claims about teacher effectiveness in light of these measures requires the collection of additional validity evidence that shows they are appropriate for this purpose (Bell, 2012).

Specification of the domain of interest is not only important to ensure the right type of validity evidence is collected, it is also necessary to ensure that system-based measures are being used and interpreted as intended. For example, a state may articulate that they are evaluating the domain of Teaching Effectiveness as measured through the constructs of Student Learning and Professional Practice. However, given the frequency with which these terms are used, to prevent individuals from making inaccurate generalizations based on personal beliefs or expectations, it is necessary to describe exactly how they are defined and operationalized within a given system.

It is also important to note that no definition of effectiveness can be "all inclusive." That is, it is not possible to measure every construct that could arguably be relevant, important, or necessary to inform the evaluation of an educator's performance. For example, while most would argue that a teacher's content knowledge of the area in which they are providing instruction clearly influences their effectiveness; most EES do not directly evaluate teacher content knowledge as a component of their system. That is, teachers are not asked to take one or more content area tests to see how well they understand the content they are expected to teach within their grade. Similarly, while some schools/districts decide to include data resulting from student perception surveys others do not. States must make decisions as to which constructs and associated

measures they believe are most important to the goals of their system. While the omission of specific constructs or measures may be problematic, it can be argued that peripheral factors not specifically targeted for evaluation are indirectly addressed through their relationship with the constructs and measures that are defined. For example, since poor content-area knowledge will likely inhibit a teacher's instructional skills, and such skills are a large focus of many observation rubrics, this factor is indirectly evaluated within most models even if it is not directly measured.

The process of sampling from the domain all factors that could be necessary to support the evaluation of educators is similar to identifying a subset of standards or objectives from a state curriculum to be the focus of an assessment used to support proficiency-based decisions about students. For a variety of reasons, both psychometric and practical, it is not possible to test the full range of knowledge and skills expected to be acquired by students within a given school year; however, a thoughtful test design developed in consideration of the goals of assessment and how results are to be used can provide-a valid, reliable estimate of student proficiency that supports the decision making process. Similarly, it is not possible to measure the full range of practices teachers are expected to demonstrate and outcomes they are expected to facilitate; however, a coherent system developed in light of a well articulated TOA can be used to provide valid and reliable estimates of teacher efficacy that support the overarching goals of the system. It is a state or district's unique perspective on the wide range of issues outlined above that leads to different system designs, and consequently the type and range of activities necessary to support validation. As a result, these factors are the focus of the framework outlined below.

The EE Validation Framework

ECD is a framework to support the design, development, and implementation of educational assessments (Mislevy, Steinberg, Almond, 2003). Within the context of ECD, assessment development occurs in a series of phases or layers, each of which informs and builds upon those adjacent to it. For example, the first phase in the ECD framework — often referred to as Domain Analysis — necessitates specification of the domain of interest in light of the defined purpose for assessment. This is followed by identification and prioritization of those components of the domain that should be the focus of assessment, the claims we want to make about students in light of assessment results, and the manner and type of evidence necessary to

support those claims. It is the layered structure of the ECD framework that provides for the creation of a comprehensive assessment argument and ultimately provides evidence to support the intended use and interpretation of assessment results (e.g., Huff & Plake, 2010 Huff, Steinberg, & Matts, 2010; Mislevy, et al, 2006).

Given the clarity and detail afforded by the use of ECD, a similar approach was taken in developing the validation framework for EES outlined in Figure 2, below.

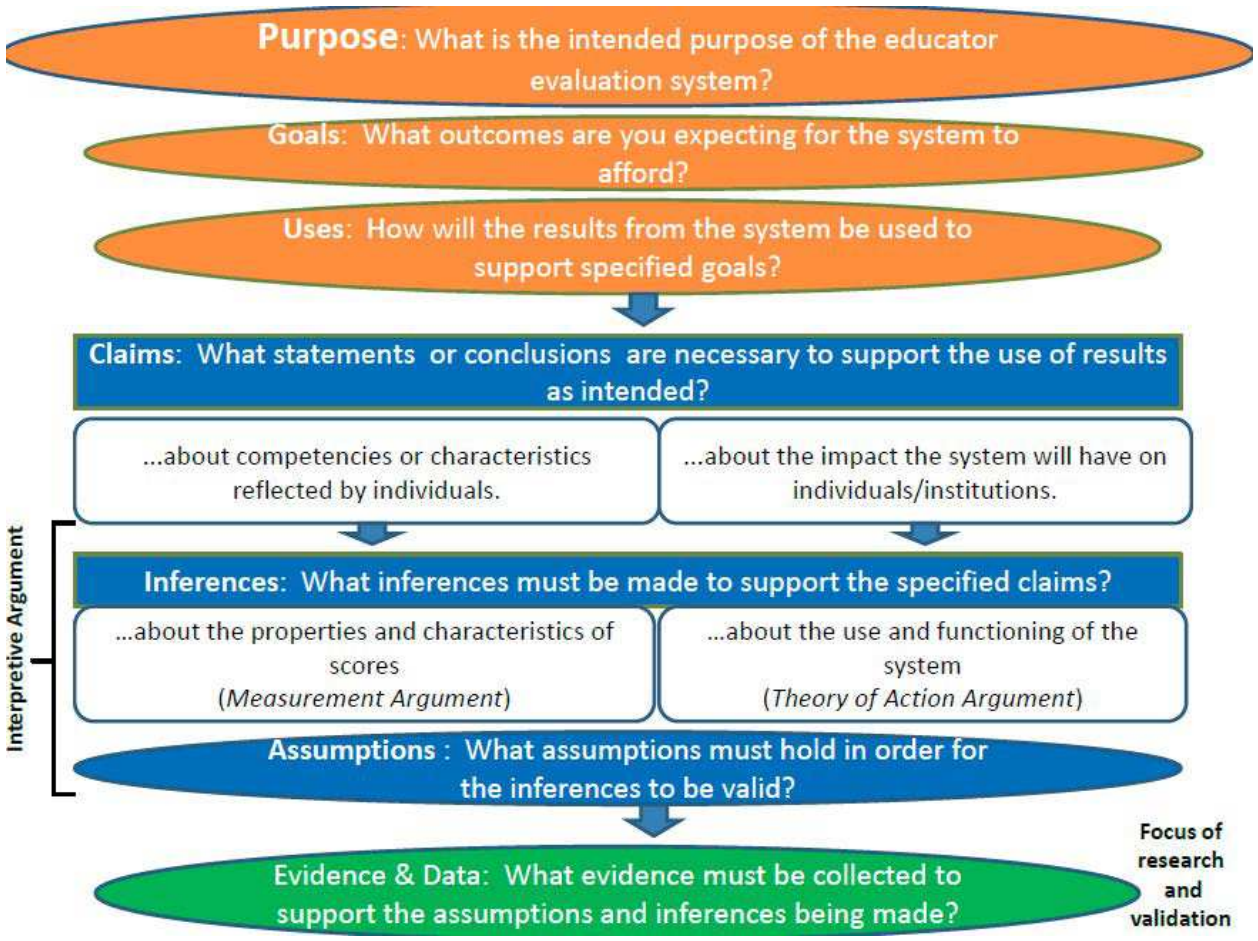

Figure 2. Structure of the Validation Framework

The validation framework, presented in the context of Figure 2 can be broken into three inter-related pieces which jointly define the TOA for the system and the evidence necessary to support it. The three pieces are differentiated by color and include: 1) the Purpose, Goals and Uses of the System; 2) the Claims and Interpretive Argument, and 3) the Evidence and Data necessary to

support validation. The framework represents an ECD-approach to validation because all elements: align to the purpose, goals and proposed uses of the system; articulate system-based claims and the evidence necessary to support them, and are explicitly stated and hierarchical (i.e., elements both build upon and inform one another). In addition, the framework promotes an iterative process of review and revision where information gained at one level of the system both informs and validates the manner in which other are specified. Each piece of the framework and its component parts are described in the sections which follow.

Purpose, Goals & Uses

The first piece of the framework includes the purpose statement, goals, and intended uses of results. The purpose statement is the overarching reason for the development of the model. It describes, in very broad terms, the driving force behind the development of the system. The Society for Human Resource Management (SHRM) differentiates three common purposes for personnel evaluation[4]: to provide management with the opportunity to evaluate employee performance and acquire information that supports administrative decisions (e.g., promotion, retention, removal); to evaluate the extent to which there is a strategic alignment between an employee's goals and strengths and the work they are doing for the organization; and to help employees develop and improve their future performance. While most EES strive to serve all three of these purposes, typically one purpose takes precedence when making design decisions.

Goals are the high-level outcomes or end results expected from the development and implementation of an EES. Goals expand upon, but are consistent with, the overall purpose of the system and provide the basis for evaluating system success. For example, when the overarching purpose of an EES is administrative, a key goal may be to establish valid, reliable measures that support the sorting selection and categorization of educators. On the other hand, when the purpose is developmental, goals typically relate to defining procedures or obtaining information that facilitates employee and system-based improvement. Goals identify the means by which the utility of a system can be evaluated by answering the question "What results do I expect to see if the system is working as intended?"

[4] Taken from: http://www.sagepub.com/upm-data/45674_8.pdf

The intended uses outline the myriad ways in which the state expects to utilize the data and information resulting from implementation. To be relevant and defensible, each intended use should align to one or more of the goals outlined for the system, as uses that are not specified and/or do not align to a system goal may not be supported by the system design. Some common uses information resulting from EESs include the following: identify high and low performing educators, support longitudinal evaluation of educator performance, highlight areas of strength and need at the educator and/or system level, target professional development activities, facilitate communication between and among educators and administrators, and support the evaluation of local programs or initiatives.

These three elements (i.e., purpose, goals, and uses) jointly dictate the requirements of the system and, therefore, drive the system design. Similar to the ECD process outlined for assessment design, it is through the consideration of these factors that a state identifies the overarching domain of interest, determines the constructs that should be evaluated in order to make inferences about that domain, and determines the outputs or measures that should be collected to quantify the extent to which the constructs are represented. In addition, at this time expectations about how the constructs underlying the system should or should not relate to one another are also expressed.

To illustrate, Figure 3 provides a hypothetical system design for State A, as described in Table 1. In this example, the state has defined the domain of interest as teaching effectiveness. For clarity, the domain is partitioned into two sub-domains – Teacher Practices and Teacher Impact. The sub-domain of Teacher Practices is conceptualized in terms of three constructs, each of which is operationalized in terms of an associated measure:Teacher Observation Rating, Educator Portfolio Rating and Student Perception Rating. Teacher Impact is conceptualized in terms of one construct, student growth, which is operationalized in terms of two measures: a VAM Rating and an SLO Rating.

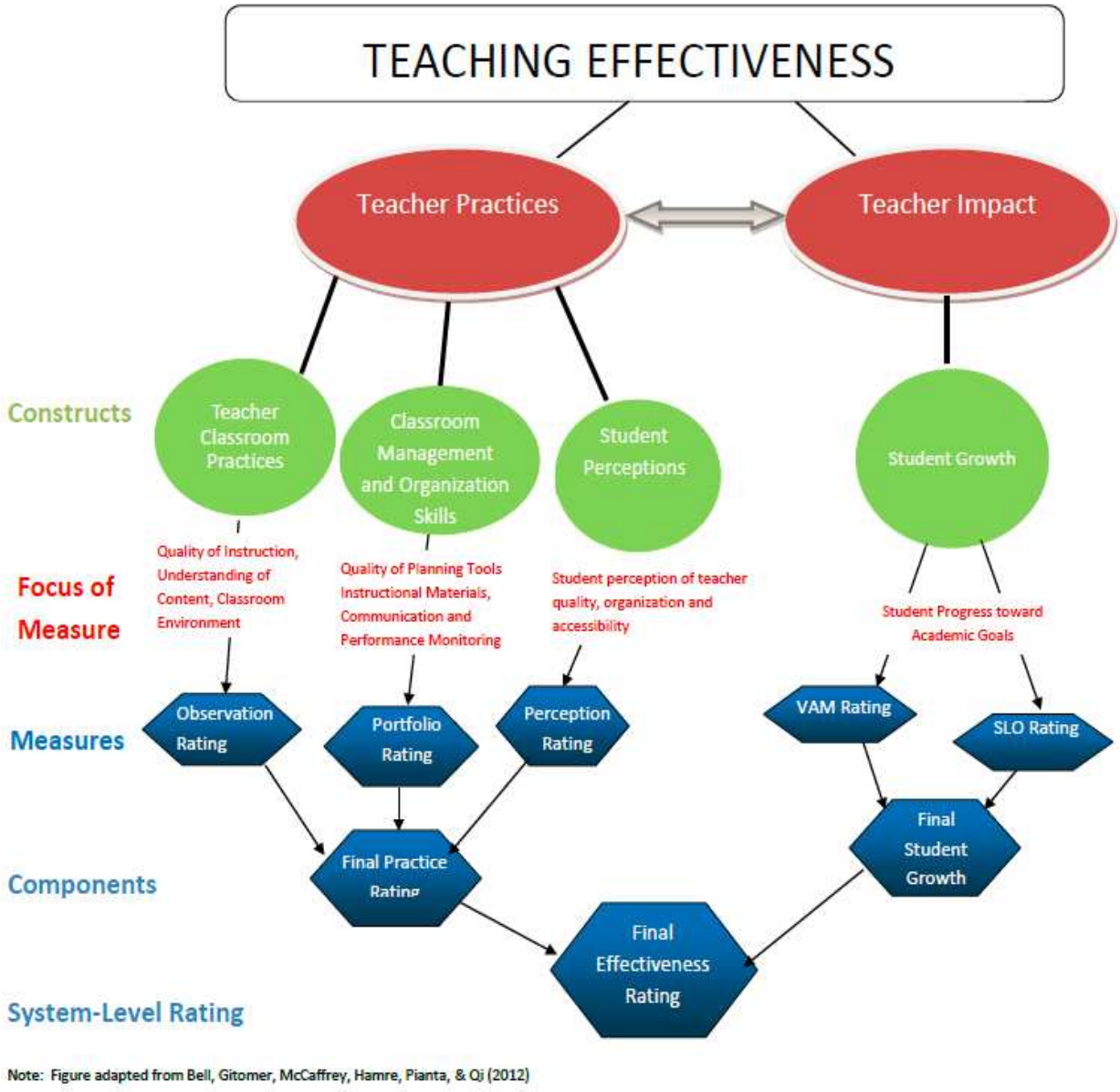

Figure 3. Summary of State A's System Design. Figure adapted from D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 1-26.

Although simplistic, a figure such as this is important to validation, because it delineates the construct each measure is intended to represent and any expected relationship among constructs. The double arrow between Teacher Practices and Teacher Impact, for example, reflects the belief that, within this system, these sub-domains are not independent and that the constructs underlying each are somehow related. While the nature of this relationship is not apparent from

the illustration, this detail is important because it suggests an additional set of analyses will be necessary to validate that the system is functioning as intended.

Claims and Interpretive Argument

The second piece of the validation framework includes claims, inferences and assumptions. Together, these elements provide for the development of a comprehensive design argument that must be supported in order to use system-based results in the manner intended. This piece of the framework focuses specifically on those elements of the TOA necessary to (a) support the specification of a comprehensive interpretive argument and (b) determine the range and type of evidence that must be collected to defend the use of system-based results.

Claims.

Claims are statements that you want to make, or conclusions which must hold true, in order to support the use of system-based results as intended. Within the context of this framework, claims are classified into one of two categories: (a) score-based claims about the competencies, characteristics and needs of educators, or (b) claims related to the impact that the evaluation system will have on different stakeholders and the mechanism by which this will occur. Score-based claims answer the question "What conclusion(s) do I want to be able to make about an educator given this score or rating in order to *use these results as intended*?" Consequently, it is important to note that there is no one right way to write a claim. The manner in which the claim associated with a system component or system-based measure is articulated (i.e., the level of specificity, area of focus, etc...) depends specifically on the use or decision the claim is intended to support. To illustrate this fact, consider the first two claims, outlined in the list of exemplar score-based claims provided in Table 2.

Table 2

Examples of score-based claims for EESs

1. Educators who obtain a high professional practice rating display the instructional skills and create a classroom environment expected from teachers.
2. Educators who obtain a low professional practice rating will benefit from targeted professional development related to instruction and/or creating a positive classroom environment.

Table 2 (cont.)

3. Educators who score low on VAM are not effective at influencing student performance as reflected by performance on the state assessment.

4. Educators who score high on VAM are effective at identifying and instructing those assessment targets upon which students are most struggling.

5. Educators who receive a final rating of Not Effective do not yet reflect the skills and competencies believed necessary to be an educator within the state.

Examples 1 and 2 represent two types of claims one might want to make with a Professional Practice score or rating. The former focuses on the use of results to classify educators relative to defined expectations, while the latter focuses on the identification of educators who will most benefit from professional development. While the measure may support both of these uses, it is the use defined and prioritized by the State relative to this measure that should characterize the focus of the claim. Claims drive the specification of inferences and influence how evidence is collected and prioritized, so articulating claims in light of intended goals and uses is crucial to the validation process.

On the other hand, impact-based claims are those that answer the question "What impact am I expecting the EES to have on different systems (e.g., schools, districts) and stakeholder groups in order to support the attainment of specified goals?" A few impact-based claims that may be associated with a given EES are provided in Table 3.

Table 3

Examples of Impact-Based Claims

The system will:

- help **teachers** improve their teaching practices.
- improve a **teachers'** ability to facilitate student learning.
- increase **teachers'** faith in the accuracy and fairness of effectiveness classifications.
- provide **teachers** and **administrators** with an increased sense of community and collaboration within and between schools.

Table 3 (cont.)

- provide **teachers** and **administrators** with greater clarity around expectations for performance.
- increase **teacher/district/public** confidence in the defensibility and accuracy of staffing decisions.
- increase **student** achievement/growth.

To support validation, in addition to the primary impact statement, the mechanism by which the system is intended to provide for that impact must also articulated. Such design statements, as we will refer to them in this paper, are necessary to help ensure that the full range of inferences and assumptions underlying a given claim will be documented and understood. For example, within our State A example, the design statements outlined in Table 4 could be presented in association with the impact claim that the system will "help teachers improve their teaching practices".

Table 4

Sample Impact Claim and Associated Design Statements

Impact Claim	**In what way will the State A's system facilitate this impact?**
The system will help **teachers** improve their teaching practices	• Observations and artifacts resulting from the evaluation process allow the evaluator to provide information to educators regarding areas of strength and weakness. • Information resulting from the student perception survey tells educators where students believe they need to improve. • The process of developing, monitoring and assessing student learning objectives facilitates good instructional and assessment practices. • Educators identified as Partially Effective or Not Effective will receive support and a detailed, specified improvement plan. • Teachers will be provided with professional development opportunities targeted at identified areas of need.

While many claims are specified and understood prior to the design phase, some emerge or result from the design process itself. Often information is obtained through piloting and preliminary analysis which informs the manner and degree to which a given claim will be supported in light

of the current system design. For example, if State A decides that they can only provide educators with an overall student perception rating, the second bullet in the list provided in Table 4 would go away. While this does not necessitate the removal of this claim, as there are many other ways in which the system is intended to support it, multiple deletions would suggest modification to the system design, or associated claim, may be necessary.

Interpretive Argument.

Once claims are specified, the inferences and assumptions necessary to support those claims must be articulated. This is often referred to as the interpretive argument. The interpretive argument articulates the conclusions one must make and the conditions which must hold in order to go from acquired data and/or information to a desired claim (Kane, 2006). Taken together, the inferences and assumptions provide specifications for the research and validation effort and clarify the nature of the evidence necessary to support the use of the system results as intended.

Consistent with the specification of claims, the validation framework distinguishes between two types of inferences: (a) inferences necessary to support the use of scores (or ratings) for making qualitative claims about educators, and (b) inferences necessary to support statements as to how/why the EES, as designed, will provide for the expected impact. This dichotomous conceptualization was inspired by an approach outlined by Bennett, Kane & Bridgeman (2011) in which they represent the Interpretive Argument in terms of two distinct, yet related parts: a measurement argument and a theory of action argument.

Measurement Argument.

Within the context of this document, the measurement argument defines how you move from data resulting from the system (i.e., student responses, scores, ratings) to claims about proficiency or competency (e.g., professional practice). The measurement argument focuses specifically on the interpretation of scores and the network of inference and assumptions that are necessary to support them (Kane, 2006). If the measurement argument is plausible, it follows that it is reasonable and appropriate to use the scores resulting from the system in support of score-based claims (Bennett, et.al, 2011).

While the specification of a comprehensive measurement argument is outside the scope of this paper[5], Table 2 provides an abbreviated example of a measurement argument for the Student Outcome Component Rating associated with State A[6]. The left side of the table outlines the inferences being made and the right side of the table details the assumptions it depends on. In this example, consistent with Kane's representation of an interpretive argument, we are making inferences related to scoring, generalization, extrapolation and a decision. Since the Student Outcome Rating is determined using a decision matrix approach that considers an educator's VAM Rating in conjunction with his/her SLO Rating the assumptions underlying each of these measures is reflected in the sample table below. In practice, a measurement argument would be outlined for each score or rating produced at each of the three levels reflected in Figure 1 (i.e., Measure, Component, and Overall Rating).

Table 5

A Sample Measurement Argument for State A's Student Outcome Rating

Inference	Assumption
Scoring Student Outcome Rating accurately reflects educator performance related to the attainment of academic targets (i.e., progress) and student growth.	S1. Scoring criteria for individual SLOs are clearly articulated. S2. Rules for calculating an overall SLO Rating are clear and appropriate. S3. SLO scoring rubrics were applied accurately and consistency. -Evaluators were adequately trained. -Evaluators used the most up to date versions of scoring and evaluation materials. S4. VAM calculations were applied accurately and consistently S5. Assumptions and scoring rules underlying VAM calculations are reasonable and appropriate. S6. SLO and VAM scoring procedures are bias free. S7. Student Outcome Decision Matrix was applied accurately. S8. Scoring rules reflected in the Student Learning Matrix are appropriate and reasonable.

[5] See Kane (2006) for a comprehensive discussion.
[6] See Bell (2012), Figure 3 for an interpretive argument specific to the validation of Professional Practice measures.

Table 5 (Cont.)

Inference	Assumption
Generalization Student Outcome Rating is representative of what would be expected across all possible observations.	G1. Conditions underlying SLO and VAM calculations are representative of that which would typically be observed. G2. The evaluator was able to adequately assess all criteria outlined for a given SLO. G3. The number of observations provide for adequate information to allow for generalization.
Extrapolation Student Outcome Rating provides information about an educator's effectiveness at facilitating student growth in achieving defined academic targets.	E1. Assessments selected/developed to evaluate attainment of SLOs are appropriate (i.e., provide for reliable scores and valid inferences related to the defined learning target). E2. SLO assessment results provide for an evaluation of student progress resulting from educator instruction. E3. SLO's reflect an appropriate and expected level of rigor for identified student populations. E4. VAM calculations reflect growth that can be attributed to the influence of the educator.. E5. SLO and VAM calculations are not influenced by extraneous factors that would seriously bias the interpretation of the Student Outcome Rating.
4. Decision Student Outcome rating supports decisions regarding ability of educator to elicit student progress and growth	D1. Educators who achieve a low student outcome rating are not likely to be successful in establishing and/or attaining academic targets.

Note: Adapted from Kane (2006)

Theory of Action (TOA) Argument.

In contrast to the measurement argument, the TOA argument describes how you go from inferences about the use, interpretation and quality of the system to claims about the expected impact of the system on stakeholders or institutions. The TOA argument focuses on the impact of system elements and the inferences and assumptions underlying the proposed mechanism by which desired impacts (e.g., modified actions, perceptions, and behaviors) will be realized. If the TOA argument is plausible, it is reasonable to assume the system will provide for the desired impact in the manner expected.

A TOA argument for the impact claim outlined in Table 4 – The system will help **teachers** improve their teaching practices – could be constructed in a variety of ways. For example, a third column could be added to Table 4 and the assumption(s) underlying each bullet in the second column could be articulated in a one-to-one fashion. A second technique would be to look over the bullets in the second column of Table 4 with the goal of identifying different types or categories of inferences. This technique is appealing because it accounts for the fact that there will be several impact claims to evaluate in practice, many of which will share similar statements about the how the system will drive change. For example providing teachers "with professional development opportunities targeted at identified areas of need" could easily be considered a key mechanism by which several of the impact-based claims outlined in Table 3 would be attained.

In addition, the identification of categories of inferences is often intuitive, as such categories typically reflect the state's beliefs regarding the overarching mechanism by which change will occur and are consistently reflected in the system design. For example, the design statements provided in Table 4 suggest that feedback, participation, and the provision of support are all necessary for the system to have the desired impact on teacher practices. The first two bullets, specifically, rely on an inference related to the provision of data and feedback to educators.

Once categories are identified, the assumptions necessary to support the array of statements associated with those categories can be articulated. Table 4 lists a few of the assumptions that must hold in order for the system to "Help teachers improve their teaching practices" in a manner consistent with that expected, as outlined in Table 4. For illustrative purposes, the assumptions have been organized relative to the three categories previously discussed.

Table 6

Abbreviated Example of a Theory of Action Argument

Inference Category	Assumptions
Feedback	• All educators are provided with feedback. • Feedback is individualized, timely, informative and useful. • Teachers understand or are given the support necessary to use provided feedback to improve knowledge/skills/practices (as appropriate).

Table 6 (Cont.)

Inference Category	Assumptions
Professional Development/ Support	• Educators take advantage of provided professional development opportunities. • Educators identified as Not Effective or Partially Effective are provided with support and an individualized improvement plan.
Active Participation in the Process	• Educators are active participants in the SLO process. • Educators and evaluators attend and contribute to all scheduled conferences.

In practice, one could imagine that a table such as this would be much longer as additional categories of inferences and associated assumptions would need to be identified to support the full array of impact claims and design statements underlying the system.

Evidence and Data

The third piece of the validation framework necessitates articulation of the evidence and data necessary to evaluate the assumptions identified within the context of the interpretive argument. It is at this point that the benefits of utilizing an ECD-based approach are most greatly realized since the full range of evidence necessary to support the utility and defensibility *of a specific EES* can be easily articulated in light of information collected at previous layers of the framework. To briefly illustrate how the specification of evidence flows from assumptions outlined in the interpretive argument, Table 7 outlines evidence that could be compiled (E), and analyses that could be conducted (A), to support the first three assumptions (S1-S3) associated with the scoring inference outlined in Table 5.

Table 7

Evidence and Analyses to Support Scoring Assumptions (S1-S3)

Assumption	Evidence/Analyses to Support Assumption
S1. Scoring Criteria for individual SLOs are clearly articulated	• (E) Feedback from evaluators regarding the perceived clarity of the SLO scoring rules and process. • (E) Feedback from evaluators regarding the ease with which the attainment of a given SLOs could be determined in light of provided scoring guidelines.

Table 7. (Cont.)

Assumption	Evidence/Analyses to Support Assumption
	• (E)Feedback from independent reviewers regarding the appropriateness of the performance category descriptors established to support the scoring of individual SLOs. • (A) Conduct a usability analysis - ask a sample of educators unrelated to the program to review and evaluate the clarity of the scoring rules associated with a set of SLOs.
S2. Rules for calculating a Final SLO Rating are clear and appropriate.	• (E) Summary of process used and stakeholders involved in determining the Final SLO Rating scoring rules. • (E) Feedback from independent reviewers regarding the appropriateness of the process used to establish a final SLO Rating.
S3. SLO scoring rules were applied accurately and consistency	• (E) Pilot data showing the extent to which evaluators provided the same scores as experienced scorers on training sets. • (E) Feedback from evaluators regarding confidence in accuracy of individual SLO scores and the overall SLO rating. • (A)Assess the degree to which two trained evaluators scoring the same set of SLOs assign similar scores. • (A) Independently calculate the overall SLO rating for a sample of educators and evaluate agreement.

A similar process could be conducted for each of the assumptions defined within the context of the TOA argument. For example, one could think about the different types of evidence that could be collected to validate the range of assumptions outlined within the inference category related to Professional Development. Such evidence could take many forms and necessitate a variety of data collection techniques, including: interviews, surveys, longitudinal analysis and research studies.

In addition to outlining evidence/data that could be used to confirm or challenge each stated assumption, activities at this layer of the framework should also include detailing (a) the process, effort and resources necessary to collect that evidence (b) what entity (i.e., state, district, school, teacher) would be responsible for its completion, and (c) the extent to which collection of that evidence would be audited or required by the state. Such information is necessary to understand the scope of the validation effort and help prioritize the wide range of activities that could occur.

From the abbreviated examples provided above it is clear that the array of evidence that could be collected to support the full range of inferences underlying an EES could be extremely large. In most cases, resources available for evaluation activities will not be sufficient to support careful examination of all of these aspects of the EES. Therefore, it is important to establish an approach for prioritizing and narrowing the validation effort so that is it as effective and efficient as possible. Such an approach should take into account not only the importance of different pieces of evidence in supporting a given set of assumptions, but also issues related to accessibility, feasibility, time and cost. Some data can be collected relatively quickly and is easily available. Other data, such as that which requires the development of a survey or structured research design will be timely and expensive. Often times there will be data/information that, once collected, serves to provide evidence in support of several different assumptions. Consequently, the importance and utility of certain types of data and analyses will vary from state to state – providing proof once again that it is the design of the system and the priorities reflected within it that reflects and dictates how a validity argument should be framed.

In the end, it will fall to States and Districts to look across the full array of evidence that could be collected, and outline the validation effort in light of prioritized goals and resources. If information has been collected and documented using the framework presented in this document, the relationship between and among elements of the system will be transparent as will the core set of evidence necessary to support key system-based claims. Consequently, states/districts will have the foundation necessary to craft appropriate short-term and long-term validation plans that best meet their goals.

Conclusion

Educator evaluation systems are being installed across the country to support informed decision making, provide useful feedback so educators can use them to inform instructional improvements, and meet federal accountability requirements. To examine whether EES are working as intended, and to defend the utility of such systems to policy-makers, tax payers and other stakeholder groups, thoughtful evaluation of these systems is necessary. This document provides a framework to support the identification and documentation of those elements necessary to establish a clear, coherent validation plan for EESs. By utilizing the principles of

evidence centered design, the framework makes the link between the overarching purpose and goals of the system and the evidence and analyses identified as necessary to support validation both rational and clear.

In addition, use of the framework serves to

- facilitate transparency;
- help stakeholders understand the rationale behind the design of the system and what it is attempting to accomplish;
- provide a common language related to educator evaluation within the context of a given system;
- highlight potential areas of incoherence in the system design;
- facilitate the identification, mitigation and evaluation of potential unintended consequences and
- support requirements related to the development and review of district-developed alternate evaluation systems (or system components).

Although only peripherally discussed within the context of this paper, the identification of potential negative consequences is critical to the specification of a comprehensive plan for validation. If system components are working as intended (e.g., they differentiate among teachers), but result in consequences that contradict the attainment of goals (e.g., high quality educators leaving the field), the design of the system and underlying theory of action will need to be revisited. Consequently, the articulation and evaluation of unintended outcomes should occur at a level of fidelity equivalent to that afforded to intended consequences.

In addition, the framework serves as a powerful tool to support state decision making regarding the manner and degree of flexibility that should be afforded to districts around the design and implementation of alternate, or aligned, EES designs. In fact, one can think about the different degrees of flexibility provided to districts as existing on a continuum that varies in terms of the extent to which a district must emulate the state-defined components of the Theory of Action. This continuum, reflected in large part by the different levels of the validation framework, is presented in Figure 4 below.

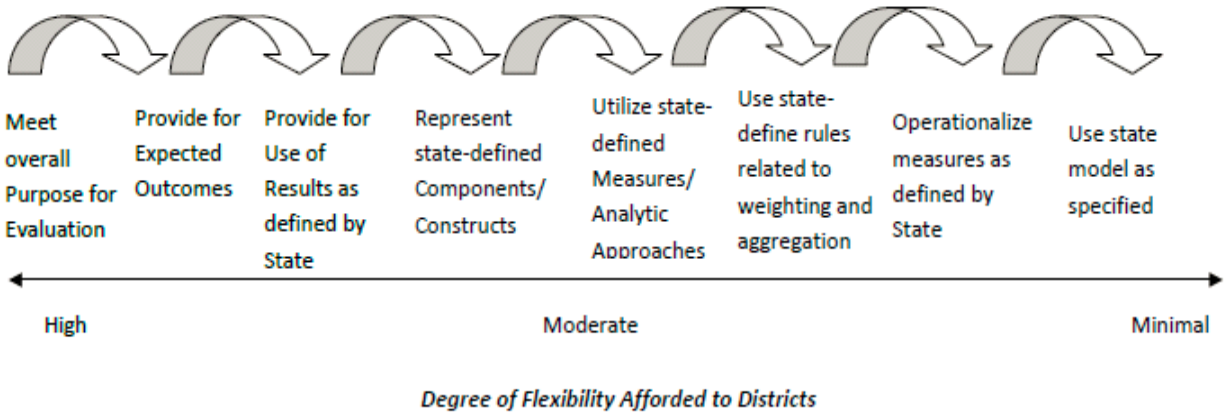

Figure 4. EE System Design Flexibility Continuum–Represented in Terms of Alignment to the State-Defined TOA.

It is important to note that as one moves from the left end to the right end of the continuum the extent to which the claims, inferences and assumptions underlying the design of the system are consistent with those defined by the state model increases, and the burden on the district (rather than the state) to collect evidence in support of these claims, inferences and assumptions decreases. Many states exist somewhere in the middle of this continuum, whereas certain elements of the TOA are prescribed, but other elements are left to the districts to defined and operationalize.

While the range of information required to support the use of a framework such as this may initially seem overwhelming, the collection of such information is critical to support the defensibility of the system, and the task is far less daunting if one has considered the need for validation early on. For most states/districts the questions posed within the first two pieces of the framework will have been addressed (if not documented), to some extent, during system design. Furthermore, if the process of validation was approached as an ongoing activity that occurs in conjunction with the design and development effort, rather than during and after implementation, much of the evidence necessary to support validation will have already been identified — and possibly collected. In these situations the framework serves more as a guide to support the documentation and rationalization of a proposed research agenda, rather than a means of generating a validation plan from scratch.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Baker, E.L. Barton P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., and Shepard, L.A., (2010). Problems with the use of student test scores to evaluate teachers (EPI Briefing paper 278). Washington, D.C.: Economic Policy Institute.

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.

Bell, C. A., Gitomer, D. H., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 1-26.

Bennett,R., Kane, M. & Bridgeman (2011). Theory of Action and Validity Argument in the Context of Through Course Summative Assessment. A paper presented at the Invitational Research Symposium on Through Course Assessment. Retrieved from http://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Bennett_Kane_Bridgeman.pdf

Burling, K. (2012). Evaluating Teachers and Principals: Developing Fair, Valid, and Reliable Systems. Retrieved from: http://educatoreffectiveness.pearsonassessments.com/

Diaz-Bilello, E., Hall, E., Marion, S., (2014). Considerations for Establishing Performance Standards for Educator Evaluation Systems. A paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

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

Goldhaber, D. and Hansen (2008). Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. CEDR Working Paper 2010-3: University of Washington.)

Goldhaber, D. & Loeb. S. Carnegie Knowledge Network, "What are the Tradeoffs Associated with Teacher Misclassification in High Stakes Personnel Decisions?" Last modified April 2013. URL = http://carnegieknowledgenetwork.org/briefs/value-added/teacher-misclassifications/

Grossman, Loeb, Cohen, Hammerness, Wycko, Boyd, and Lankford (2010). Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added. NBER Working Paper 16015.

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

Huff, K., Plake, B., (2010). Evidence Centered Design in Practice, *Applied Measurement in Education*. Vol 23. Iss. 4; 307-309.

Huff, K., Steinberg, L., Matts, T., (2010) The Promises and Challenges of Implementing Evidence-Centered Design in Large-Scale Assessment. *Applied Measurement in Education*: 23 (4); 310-324

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement*, 4th ed (pp. 17-64). Westport, CT: Praeger

Marion, S. (2010) Developing a Theory of Action: A Foundation of the NIA Response, Retrieved from http://www.nciea.org/

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

Mislevy, R. J. and Haertel, G. D. (2006), Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25: 6–20. doi: 10.1111/j.1745-3992.2006.00075.x

Mislevy, R.J., Steinberg, L.S., & Almond, R.A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspective*s, 1, 3-67.

Sheppard, L., (2012). Evaluating the Use of Tests to Measure Teacher Effectiveness: Validity as a Theory-of-Action Framework, A paper presented at the annual meeting of the National Council on Measurement in Education, Van Couver, British Columbia.

Weisberg, D., Sexton,S., Mulher, J.,and Keeling, D."The Widget Effect." The New Teacher Project, 2009. http://widgeteffect.org/downloads/TheWidgetEffect.pdf..