State and Local Efforts to Investigate the Validity and Reliability of Scores from Teacher Evaluation Systems

Corinne Herlihy, Ezra Karger, Cynthia Pollard, Heather C. Hill,
Matthew A. Kraft, Megan Williams, Sara Howard

Structured Abstract

Context: In the past two years, states have implemented sweeping reforms to their teacher evaluation systems in response to Race to the Top legislation and, more recently, NCLB waivers. With these new systems, policy-makers hope to make teacher evaluation both more rigorous and more grounded in specific job performance domains such as teaching quality and contributions to student outcomes. Attaching high stakes to teacher scores has prompted an increased focus on the reliability and validity of these scores. Teachers unions have expressed strong concerns about the reliability and validity of using student achievement data to evaluate teachers and the potential for subjective ratings by classroom observers to be biased. The legislation enacted by many states also requires scores derived from teacher observations and the overall systems of teacher evaluation to be valid and reliable.

Focus of the study: In this paper, we explore how state education officials and their district and local partners plan to implement and evaluate their teacher evaluation systems, focusing in particular on states' efforts to investigate the reliability and validity of scores emerging from the observational component of these systems.

Research design: Through a document analysis and interviews with state education officials, we explore several issues that arise in observational systems, including the overall generalizability of teacher scores, the training, certification, and reliability of observers, and specifications regarding the sampling and number of lessons observed per teacher.

Findings: Respondents' reports suggest that states are attending to the reliability and validity of scores, but inconsistently; in only a few states does there appear to be a coherent strategy regarding reliability and validity in place.

Conclusions: There remain a variety of system design and implementation decisions that states can optimize to increase the reliability and validity of their teacher evaluation scores. While a state may engage in auditing scores, for instance, it may miss the gains to reliability and validity that would accrue from periodic rater retraining and recertification, a stiff program of rater monitoring, and the use of multiple raters per teacher. Most troublesome are decisions about which and how many lessons to sample, which are either mandated legislatively, result from practical concerns or negotiations between stakeholders, or, at best case, rest on broad research not directly related to the state context. This suggests that states should more actively investigate the number of lessons and lesson sampling designs required to yield high-quality scores.

Executive Summary

Context:

In the past two years, states have implemented sweeping reforms to their teacher evaluation systems in response to Race to the Top legislation and, more recently, waivers of The No Child Left Behind Act of 2001. With these new systems, policy-makers hope to make teacher evaluation both more rigorous and more grounded in specific job performance domains such as teaching quality and contributions to student outcomes. Each of these new teacher evaluation systems produces overall performance scores for individual teachers that are derived from multiple sources of data, including classroom observation systems. What these scores mean, and how reliably they measure differences in teacher or teaching quality, is an open question.

In some states, new legislation also attached important consequences to the performance evaluation scores teachers receive. For example, teachers who receive excellent ratings may receive financial bonuses, salary increases, non-probationary status, or tenure. Teachers judged as performing poorly may be denied pay raises or tenure, enrolled in mandatory assistance or remediation plans, or terminated. Attaching high stakes to teacher scores has prompted an increased focus on the reliability and validity of these scores. Teachers unions have expressed strong concerns about the reliability and validity of using student achievement data to evaluate teachers and the potential for subjective ratings by classroom observers to be biased. The legislation enacted by many states also requires scores derived from teacher observations and the overall systems of teacher evaluation to be valid and reliable.

Focus, Design, and Sample:

In this paper, the authors explore how state education officials and their district and local partners plan to implement and evaluate their teacher evaluation systems, focusing in particular on states' efforts to investigate the reliability and validity of scores emerging from the observational component of these systems. To this end, we began with a document analysis of teacher evaluation legislation and guidelines in 17 states. We then conducted a series of interviews with officials representing 12 states, asking about current concerns, efforts, and issues surrounding the production of high-quality teacher scores. We focused in particular on areas known to be of concern in the generation of high-quality observational scores, including the choice of the observational instrument, rater training and certification, and the number of lessons evaluated per teacher per year.

The sample of 17 states used in this study satisfy four requirements: each state received a Race to the Top grant or a No Child Left Behind waiver before July 1, 2012; conducted a pilot-test of its new teacher evaluation system in a subset of schools or districts during or before the 2012-2013 school year; had statutory language describing a teacher evaluation system which satisfied the requirements of its Race to the Top grant or No Child Left Behind Waiver; and did not have any pending legislation, as of July 1, 2012, which would substantially change the statutory basis for

the state's teacher evaluation system. The latter two requirements removed six states from an original sample of 23 states which satisfied the first two conditions.

We requested interviews from individuals in each of the 17 sampled state departments of education and made repeated interview requests if no response was received. We spoke anonymously with 13 people from 12 states with each interview lasting approximately 45 minutes. The most common titles of our 13 interviewees were director, coordinator, or executive officer of the state's efforts to implement the new teacher evaluation system, but interviewees' positions ranged from researcher to state superintendent. The interviews took place in August and September 2012 as most states began to implement newly legislated teacher evaluations.

Findings and Conclusions:

Teacher evaluation systems have undergone marked changes in a very short amount of time. The majority of states we studied are currently piloting new systems or in the beginning stages of full implementation. Despite significant federal funding for these efforts, there remain significant resource constraints, most often felt at the district level where the implementation costs are largely born. These constraints will undoubtedly affect the validity and reliability of the scores produced by the new teacher evaluation systems. However, there remain a variety of system design and implementation decisions that states can optimize to increase the reliability and validity of their teacher evaluation scores even within these constraints.

Although many states have adopted one or two best practices, these seldom occur as a coordinated program of inquiry and action to achieve reliable and valid scores. While a state may engage in auditing scores, for instance, it may miss the gains to reliability and validity that would accrue from periodic rater retraining and recertification, a stiff program of rater monitoring, and the use of multiple raters per teacher. Most troublesome are decisions about which and how many lessons to sample, which are either mandated legislatively, result from practical concerns or negotiations between stakeholders, or, at best case, rest on broad research not directly related to the state context and instrument. This suggests that states should more actively investigate the number of lessons and lesson sampling designs required to yield high-quality scores.

The lack of a coordinated program may also have a large impact in the area of consequential validity, in other words, how schools, teachers, and children experience the system. Although many state respondents placed consequential validity – typically in the form of stakeholder opinion – high on their list of criteria for policy success, few described a concrete program of research that would study policy effects. We can imagine two phenomena particularly worth tracing from this point forward: the rate of incorrect decisions (e.g., tenuring poor teachers, dismissing excellent teachers), and the overall impact of the new teacher evaluation systems on teacher recruitment and attrition. As states and districts seek to recruit the best and brightest into teaching, an evaluation system that is perceived to accurately recognize and reward talent may

improve recruitment, just as one that is perceived as being arbitrary and unfair may push candidates towards other career options.

Finally, current reforms in teacher evaluation have potential consequences for a wider set of system-level features, such as teacher professional development initiatives and the resources invested in these efforts, principal training and recruitment as instructional leaders, as well as the day-to-day practice of teaching itself. Whether these reforms positively or negatively affect these features depends, in some part, on the reliability and validity of scores, as well as the design of the accountability system as a whole. Getting the measurement of teaching right, in this view, is critical to improving school and student outcomes.

State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems

Corinne Herlihy, Ezra Karger, Cynthia Pollard, Heather C. Hill,
Matthew A. Kraft, Megan Williams, Sara Howard

Introduction

In the past two years, states have implemented sweeping reforms to their teacher evaluation systems in response to Race to the Top legislation and, more recently, waivers of The No Child Left Behind Act of 2001. With these new systems, policy-makers hope to make teacher evaluation both more rigorous and more grounded in specific job performance domains such as teaching quality and contributions to student outcomes. Most states hope to achieve these aims by designing systems that accurately capture the "true" distribution of teaching skill and by requiring that these systems incorporate student performance into the evaluation score. As of July 1, 2012, 23 states had undertaken or promised reforms to their teacher evaluation systems after receiving a Race to the Top grant or a No Child Left Behind Waiver, with a majority of states having implemented full-scale or pilot evaluation systems before or during the 2012-2013 school year.

Each of these new teacher evaluation systems produces overall performance scores for individual teachers that are derived from multiple sources of data, including classroom observation systems. What these scores mean, and how reliably they measure differences in teacher or teaching quality, is an open question. Typically, authors of conventional assessments, such as student assessments or teacher certification tests, investigate and publish information about the characteristics of test scores, including their reliability and validity. However it is unclear

whether states intend and have the capacity to conduct similar analyses of this next generation of teacher evaluation systems.

In this paper, the authors explore how state education officials and their district and local partners plan to implement and evaluate their teacher evaluation systems, focusing in particular on states' efforts to investigate the reliability and validity of scores emerging from the observational component of these systems. To this end, we began with a document analysis of teacher evaluation legislation and guidelines and then conducted a series of interviews with officials representing 12 states, asking about current concerns, efforts, and issues surrounding the production of high-quality teacher scores. We focused in particular on areas known to be of concern in the generation of high-quality observational scores, including the choice of the observational instrument, rater training and certification, and the number of lessons evaluated per teacher per year (Bell et al, 2012).

Literature Review

The Obama Administration and its Secretary of Education, Arne Duncan, have leveraged federal education funding and waivers in ways previously unseen in education politics. At Obama's behest, Congress authorized $4.35 billion in funding from the American Reinvestment and Recovery Act (2009) for a competitive grant program called Race to the Top (RTTT) to motivate states and districts to create plans for improving their education systems. The largest competitive grant program ever sponsored by the federal government (U.S. Department of Education, 2009), Race to the Top identified four broad reform initiatives that states had to address in their grant applications:

(1) recruiting, preparing, retaining and rewarding effective teachers and principals (especially in high need areas and subjects);
(2) adopting standards and assessments that are aligned and adequately prepare students to succeed in post-secondary education and their careers, and to compete in the global economy;
(3) building data systems that measure student growth and can provide information to teachers and principals about how they can improve instruction;
(4) turning around the lowest-performing schools.
(U.S. Department of Education, 2009)

Department of Education envisioned the implementation of a new generation of rigorous teacher evaluation systems as the core of its efforts to build a more effective teaching force. This prioritization was evident in the grant evaluation criteria, which heavily weighted[1] proposed efforts to evaluate teachers using multiple measures that included student performance. Forty four states plus the District of Columbia applied in the first RTTT round in 2010, with Delaware and Tennessee winning funds in the amounts of $100 and $500 million respectively. In the second round, the Department of Education awarded grants, which ranged from $75 to $700 million, to 10 states. Another $200 million was divided among seven states in a third and final round.

The Obama administration has also influenced states' education reform efforts by offering Elementary and Secondary Education Act (ESEA) Flexibility or No Child Left Behind (NCLB) waivers. Announced in 2011, ESEA Flexibility grants states relief from some of the key requirements of NCLB, including the mandate that 100% of students score "proficient" in English Language Arts and Mathematics by the end of the 2013-14 school year. In exchange for this flexibility, states were required to demonstrate in their waiver requests that they:

(1) had college- and career-ready expectations for all students;
(2) had developed, and have a high-quality plan to implement, a system of differentiated recognition, accountability, and support for all Title I districts and schools in the State;
(3) were committed to developing, adopting, piloting, and implementing teacher and principal evaluation and support systems that support student achievement; and
(4) had provided an assurance that they will evaluate and, based on that evaluation, revise its administrative requirements to reduce duplication and unnecessary burden on districts and schools. (U.S. Department of Education, 2012)

By September 2012, 44 states and the District of Columbia had applied for waivers; thirty-three states and the District of Columbia had been approved for waivers; and eleven states in addition to the Bureau of Indian Education and Puerto Rico had outstanding requests for waivers.[2]

States participating in Race to the Top and ESEA Flexibility now must deliver on their commitments to create and implement new teacher evaluation systems. Many states have passed legislation that establishes the criteria, data system infrastructure, and data collection processes necessary for such large-scale systems. In some states, new legislation also attached important consequences to the performance evaluation scores teachers receive. For example, teachers who receive excellent ratings may receive financial bonuses, salary increases, non-probationary status, or tenure. Teachers judged as performing poorly may be denied pay raises or tenure, enrolled in mandatory assistance or remediation plans, or terminated.

Attaching high stakes to teacher scores has prompted an increased focus on the reliability and validity of these scores. Teachers unions have expressed strong concerns about the reliability and validity of using student achievement data to evaluate teachers and the potential for subjective ratings by classroom observers to be biased (Heitin, 2012; NEA, 2011). The legislation enacted by many states also requires scores derived from teacher observations and the overall systems of

teacher evaluation to be valid and reliable. For examples of states with statutory language referring to validity and reliability, see Appendix 1.

Just what states mean by these references is unclear, not only because most legislation fails to specify criteria for validity and reliability but also because the definitions for reliability and validity have varied both historically (Kane, 2006) and across fields. In this paper we use the definitions of valid and reliable that are traditionally found in the assessment and research community; although this community has recently moved toward incorporating both considerations within the validity argument approach (see AERA, 1999; Kane, 2002; 2006) we describe them separately here, in recognition of the fact that taking a unified approach may be unrealistic for practitioners with limited time and resources. Further, the unified approach often entails specific investigations that look much like those we and state leaders describe below.

Validity refers to whether scores from an assessment do in fact represent the underlying trait – in this case, teaching quality –which they intend to capture. Historically, researchers have investigated validity in several ways. Criterion validity, for instance, focuses on the relationship between scores from an assessment and performance in the arena the assessment is supposed to predict; SAT scores, in this tradition, may be correlated with college performance to understand the relationship between the predictor and eventual performance. In the area of teacher quality, however, it is difficult to determine which outcomes should serve as the criteria, as outcomes of teaching – college attendance, earnings – are not typically available until well after any evaluation period. State policy-makers may then turn to other methods for ascertaining score validity, including assessing the "face validity" of a tool–for instance, the extent to which experts would agree that an instrument represents the domain of "teaching." State policy-makers may also perform factor analyses or other construct identification procedures with data, for instance

demonstrating that items from an observational assessment cohere into theoretically expected dimensions. States may also investigate validity by specifying constructs that should be theoretically correlated and uncorrelated with teaching quality, measuring these constructs, and then testing to see whether the theoretical predictions are true. For example, while teachers' knowledge should be correlated with teaching quality scores, students' demographic characteristics should not be correlated with teaching quality scores, at least in an ideal system. Researchers have started to examine these issues for several major observation instruments (Bell et al, 2012; Hill et al., 2012; Martinez et al., 2012). Finally, researchers have proposed that validity inquiries should include an examination of the decisions and actions based on scores – searching for both intended and unintended consequences. This type of investigation into consequential validity can exist at the level of individual decisions or at the system level, for instance when the implementation of high-stakes testing leads to teaching to the test or other methods of score inflation, which would then prompt state policy-makers to revise their view of the validity of the instrument (Koretz , 2008).

Reliability refers to whether an assessment produces consistent scores. For instance, the reliability of commercially produced bathroom scales is typically quite high; purchase any two bathroom scales and they are likely to return the same weight for an individual. Using a scale in different contexts – a bathroom, a school, or a gym – will not change results from the instrument. In the area of observationally-based assessments of practice, however, the situation is more complex. Observers–in this case principals or other raters–judge the quality of teaching; and the harshness of judgment may vary across individuals or over time. For this reason, those developing or implementing classroom observation instruments may calculate inter-rater reliability or rater agreement with a "gold standard," which typically involves lessons scored by

master raters. Teachers' practice and classroom climate may also vary between lessons, leading to questions about which and how many lessons to sample in order to arrive at a stable teacher score. Contexts also vary: the content of the lesson may affect scores on an instrument, and teachers may also teach different groups of students over time, again influencing the stability of their score. In many cases (e.g., Bell et al., 2012; Hill, Charalambous & Kraft, 2012; Shavelson & Webb, 1991), observational data suggests significant score instability due to raters and lessons. Because of concern over score variability due to both these factors, instrument developers will often perform generalizability and decision studies, or studies that decompose variance in teacher scores and estimate reliability under different scenarios, in order to ascertain an optimal number of lessons and raters needed to generate reliable teacher scores with their instrument.

The intention of many states to use evaluation scores for consequential decisions makes the reliability and validity of these scores of central importance to individual teachers and to the education system as whole. However, standards of acceptable levels of validity and reliability for these measures are often absent or unclear in the state guidelines. A survey of standards for reliability and validity in other professions provide little assistance; there exist very few industries, such as manufacturing and investment banking, where employee evaluations can be based on objective measures of productivity such as production output or investment returns. In most labor sectors, performance reviews are based on the subjective assessments of an employee's direct supervisor, yet these are often seen as flawed indicators. For example, a recent survey conducted by the Society for Human Resource Management revealed that only 55% of human resource professionals agreed that "annual performance reviews are an accurate appraisal for employees' work" (SHRM, 2012). It may be that this is acceptable in private industry

because the goals of performance reviews are to provide feedback, identify areas for development, and inform compensation decisions. Since it is recognized that performance reviews can be subjective and managers may have different standards for performance, such reviews may trigger follow-up or a probationary action before high-stakes decisions, like termination, are made. Whether districts and states will be required to meet similar standards will undoubtedly be decided by future court cases, and many may choose to adopt policies including probationary action before final employment decisions are made until they learn more about the reliability and validity of their systems.

Beyond legal ramifications, there are reasons for states to care about the reliability and validity of scores from new teacher evaluation systems. If scores are perceived to be unreliable or biased, new evaluation systems may demoralize the teaching workforce, increase attrition, and also encourage teachers to "shop" for schools where achieving an acceptable evaluation score is easier. Such a situation may deter talented individuals from entering the teaching pool, or present an incentive structure that draws teachers away from schools with high standards. Finally, if scores assigned to teachers are inaccurate, teachers and their supervisors will not be able to focus professional development and learning opportunities where they are needed–both in terms of identifying specific teachers in need of growth, and also in terms of correctly identifying specific areas of growth for individual teachers. Because workforce development is a major goal for many states and policy-makers involved with redesigning teacher evaluation, accurate diagnostic information is critically needed.

Sample and Methods

To investigate how states are responding to either legislative or other mandates to investigate score reliability, we conducted a series of interviews in the summer and fall of 2012. The sample of 17 states[3] used in this study satisfy four requirements: each state received a Race to the Top grant or a No Child Left Behind waiver before July 1, 2012; conducted a pilot-test of its new teacher evaluation system in a subset of schools or districts during or before the 2012-2013 school year; had statutory language describing a teacher evaluation system which satisfied the requirements of its Race to the Top grant or No Child Left Behind Waiver; and did not have any pending legislation, as of July 1, 2012, which would substantially change the statutory basis for the state's teacher evaluation system. The latter two requirements removed six states[4] from an original sample of 23 states which satisfied the first two conditions.

We collected and analyzed a variety of documents to answer basic questions about each state's legislative requirements and guidance regarding their teacher evaluation system. These documents included legislation, government guidelines, Race to the Top applications, No Child Left Behind flexibility requests, and other artifacts, which we used to create a spreadsheet displaying key system characteristics, such as the minimum number of observations for veteran and novice teachers and specified consequences for teachers falling into the poor and excellent categories of their states teacher evaluation systems. Data from this spreadsheet were then reduced into Exhibit 1, which lists system characteristics.

These data helped inform our construction of a 25-question interview protocol. The protocol included sections covering state policy-makers' definitions of reliability and validity, the type of efforts they are engaged in or anticipate engaging in to determine score reliability and validity, and then separate sub-sections about the selection of observational instruments, and the training of raters. We requested interviews from individuals in each of the 17 sampled state departments

of education and made repeated interview requests if no response was received. We spoke anonymously with 13 people from 12 states with each interview lasting approximately 45 minutes. The most common titles of our 13 interviewees were director, coordinator, or executive officer of the state's efforts to implement the new teacher evaluation system, but interviewees' positions ranged from researcher to state superintendent. The interviews took place in August and September 2012 as most states began to implement newly legislated teacher evaluations.

Responses from interviews were coded thematically using the qualitative data analysis package ATLAS.ti to categorize statements about validity and reliability. We began by attaching text directly to the question that was asked so that answers to specific questions (e.g., efforts to ensure reliable and valid scores) could be directly compared. We then read these responses and compared notes about themes. Results from these discussions are summarized in paragraphs presented throughout this paper. We refer to data from the analysis of publicly available documents by identifying specific states with particular policy stances; however, to maintain respondent confidentiality, we describe responses from the interviews without identifying which state an interviewee represents.

Findings

The findings are based on both the document review and the analysis of interviews with states officials. The findings are grouped into three sections: an overview of the systems being implemented in states; state policymakers' general thoughts on the reliability and validity of their system; and system design, including designating and training evaluators, selecting or building instruments and sampling lessons. Each section is organized around the questions posed to subjects.

Overview of New Teacher Evaluation Systems

As per requirements of both RTTT and NCLB, each of the 17 originally sampled states revised their teacher evaluation system; these revisions included increasing the rigor of and in some cases standardizing teacher observation instruments. An inspection of publicly available documents reveals that two states[5] adopted a state-wide classroom observation instrument and fifteen states[6] allowed districts to select, adapt or develop their own instrument, often from a list of approved instruments. As these figures suggest, some states face substantial amounts of work in developing, piloting, supporting, and validating either new instruments or significant adaptations to existing instruments. In other cases, states have been able to purchase an instrument, training, and support – but still face challenges regarding ensuring the instrument is used properly in a local context.

As per the intent of the original legislation and waivers, scores from classroom observations and student performance metrics will contribute to a total teacher score, which will be used to make significant decisions regarding school staffing and professional development. Based on our review of legislation, all states except Arizona (16 out of 17) specify consequences of poor performance on a teacher evaluation. These consequences include teachers having their salary frozen, being required to participate in teacher assistance or remediation programs, and being dismissed or terminated. All but one state (CT) also specify some consequence for excellent performance, ranging from bonuses and salary increases to tenure, and fewer observations in the future.

Interviews with state respondents shed more light on state priorities in the use of scores. For seven states in our sample, respondents answered a generic question about intended use of scores

by talking exclusively or nearly exclusively about targeted professional development. For instance, an official from one state reported: "Well, we're hoping mostly for professional development and support. We're hoping that the evaluation is designed to give very specific and actionable feedback to teachers." Another state official reported "[the] intended purpose with our evaluation system is not to call people out and say, 'Why are you such a bad teacher?' The purpose behind it is really to help teachers that are struggling to be better teachers, so that they don't leave the profession after a year or two simply because they didn't have that ongoing support and professional development." In one state, an official expressed disbelief that the system would be used to dismiss any teachers. By contrast, interviews with two other states revealed a focus on designating and ensuring consequences for poorly performing teachers. As one interviewee described, "So we're talking employment decisions. Two years of ineffective teaching means that a teacher shall not be reemployed." In another four states' interviews, responses were mixed between professional development goals and negative consequences.

Viewed in light of the historic system for producing and using teacher evaluation scores, these are more than modest changes. More standardized and, in many policy-makers' views, more rigorous instruments are being put in place, and scores will be used to make consequential decisions for teachers in many locations. Perhaps unsurprisingly, legislation in seven states requires inquiry into the validity and reliability of scores derived from these observational systems, and another five require inquiry into the validity and reliability of teachers' total scores, which include both the observation component and the student outcomes components. However, only three of these 17 states require a public report on aspects of reliability and validity of overall teacher scores, and another five appear, based on legislation alone, to place little emphasis on ascertaining the characteristics of scores. To investigate this issue further, we turn to

state policy-makers' views of and reports on efforts to monitor the reliability and validity of scores.

State Policy Makers' View: Defining Reliability and Validity

The questions in this section were intended to elicit policymakers' thoughts about the meaning of reliability and validity, either in terms of a broad definition for these terms or the standards that would need to be met for teachers' scores to be regarded as reliable and valid. We viewed policy-makers' definitions of reliability and validity as key conditioning agents for their efforts to investigate these issues, which we subsequently asked them to talk about. Taken together, responses to these questions describe the ways in which state officials are investigating the properties of teacher scores.

In your opinion, what would it mean for teachers' scores to be reliable and valid?

Of nine states that answered this question, only one answered with a broad definition of validity ("[scores] capture the effectiveness of the teacher under each of the four standards of practice.") Other states discussed specific indicators of reliability and validity, including several criteria from the measurement literature. For instance, three state officials said that they would expect to see congruence between observation and student-assessment-based metrics. Two respondents focused on elements of the system that would improve validity (a distribution of scores that better reflects reality) and reliability (raters who are more grounded in benchmarks and evidence).

However, in answering this question, the majority of policy-makers referred not to definitions or technical requirements, but instead to the broader effects of the new teacher evaluation system, often described in vague terms. For example, one state reports that they would view "trust" among stakeholders as an indicator of score reliability and validity; another state official said that

if they knew the system as a whole was affecting "what it needs to affect," then there would be evidence of validity and reliability; a third state said that if scores were used to help educators "grow and continue to grow their practice," this would show evidence of reliability and validity. Finally, one state referenced a kind of face validity by suggesting a survey of participants to ask if they thought the new system was working or not. These comments suggest that policy-makers prioritize issues related to consequential validity in their assessment, and prioritize consequences that occur at the level of the system rather than individuals, arguably at the expense of more precise ideas about scores representing teacher effectiveness or having specific desired properties.

Finally, our review of interview transcripts suggests that most states only discussed one indicator of reliability and validity, even though past research suggests that as applied to observational instruments, these topics contain many potential avenues for investigation, including construct and criterion validity. In addition, the single indicator discussed by each state varied widely across states, suggesting that the definitions for reliability and validity as applied to new teacher evaluation systems is highly context-specific. We will revisit this issue again below.

What efforts are underway to ensure validity and reliability of the scores? In general, what would you like to see in order to assure reliable and valid scores?

Officials from each of the twelve states responded to these questions, with answers varying widely across states. Four respondents described efforts to ensure evaluators are meeting standards through enhanced training. For instance, in response to the first question, one state respondent noted "One of the first pieces right now is beginning with training. We want to make sure that every evaluator is well trained and can implement the model with fidelity." Another

state also mentioned periodic recertification of raters, a practice common in the research world to guard against rater "drift" from anchor scores. We will return to the theme of training below, where we discuss results from a question aimed at discerning states' training plans.

States also described several types of empirical investigations, either finished or planned. Three states have plans to correlate value added measures or VAM with observation scores, and to use this as evidence for validity; several states mentioned other analyses, including an investigation of inter-rater reliability and factor analyses of data; and three states discussed a planned or completed study of validity and reliability, but did not provide specifics. Several states commented on the fact that only limited resources are available for these types of activities.

States also described several planned auditing or monitoring techniques. One state plans to have state observers co-rate lessons with trained and credentialed evaluators, typically principals, and compare scores. In four states there are plans to audit scores—for instance, to identify schools where value-added is low but there are many teachers rated as proficient or above based on classroom observations. As one state official said,

> Then, when we get back around to monitoring, we could potentially publish a chart that says, for example, here's the mean growth percentile of teachers who were rated effective in District A and the mean student growth percentile of teachers who were rated effective in District B and guess what? They're different. Now if we simply say, "Guess what? They're different," that's going to potentially people asking the question, "Gee, in District B, is that what we really meant by effective, or did we mean something a little more robust than that?

One state noted that it plans to prioritize investigations of low-scoring teachers, to ensure that decisions made about those teachers are valid, and two states mentioned getting feedback from stakeholders as an important part of the validation process.

Taken together, there is evidence that states are attending to many of the issues involved in establishing reliability and validity of observational assessment systems. These issues–from construct identification methods such as factor analyses to inter-rater reliability and predictive correlations are well-established as major indicators of the properties of an assessment or assessment system (Bell et al., 2012). However, in the interview data, it was rare to see a state attending to more than a handful of these issues at a time, and some states reported attending to only one. Further, several major topics in validity research were omitted, including for most states the calculation of inter-rater agreement rates, and for all states evidence suggesting that the lessons sampled accurately represent teachers' practice.

System Design: Evaluators, Instruments and Sampling Lessons

How are raters selected, trained and monitored?

Answers to questions about how raters are selected, trained, certified, and monitored can help illuminate the extent of efforts to build valid and reliable teacher observation system. In general, research on classroom observation instruments suggests that differences in raters' use of instruments, or raters use of instruments for particular lessons, explains a fair amount of variability in teacher scores. Hill et al. (2012), for example, found that 5-30% of this teacher score variability, depending on the domain being measured, could be explained by raters. For this reason, many instrument developers recommend extended training and rigorous certification procedures, such as an examination and periodic retraining and recertification. Many also put in place monitoring procedures, such as weekly or monthly calibrations. Finally, some research-based uses of classroom observation instruments require two independent observers, either

distributed across a given teacher's lesson or for each lesson (Bell et al., 2012; Hill, Charalambous, & Kraft, 2012).

In line with procedures for conducting classroom observation in research, all states we studied require raters to participate in some type of training or professional development, although the providers and focus of such training vary widely across states. The responsibility of providing rater training varied by state. In some states, respondents reported that a tradition of local control and limited state resources led to decentralization; in others, training responsibilities rested with state departments of education or third-party contractors. Almost all states in our sample described the purpose of this training as familiarizing raters with the elements of the overall evaluation system and the specific observational instrument that will be used, including how to collect and document evidence during the evaluation process. Seven states also described how training would attend to issues of "inter-rater reliability" or rater "calibration" and "alignment." Six states also noted that improving the ability of raters to provide feedback will be a specific goal for training. For example, Florida's legislation calls for the creation of training modules on how to provide "specific, actionable, and timely feedback," while trainings in Georgia will review "best practices for providing ongoing and end-of-year feedback to teachers."

Our document analysis and interviews with state officials indicate that it is principals who will attend these trainings and serve as the primary raters in the vast majority of states. Although several states use broader terms such as "Administrators" or "Instructional Leaders," which allow for some flexibility in rater selection at the local level, most appeared to default to the current system in which principals often serve as the sole rater. Of the 17 states in our sample, only North Carolina, Indiana, and Maryland require or recommend that a second rater also evaluate teachers. North Carolina's law stipulates that probationary teachers must be observed

once a year by a peer evaluator, while Indiana recommends that independent contractors use multiple raters when evaluating teachers. Maryland requires that "an evaluation report that evaluates a teacher as ineffective shall include at least one observation by an individual other than the immediate supervisor." Yet apart from these three states with mixed systems, principals shoulder most of the burden. In some cases, respondents said that the issue of developing a peer observation system had been raised, but that, ultimately, union regulations prevented this option from becoming a reality.

Relying on principals, as most states do, has the benefit of placing a large component of teacher evaluation on teachers' direct supervisor—similar to what occurs in other sectors of the labor market. However, it also raises three challenges in practice. First, many teachers (and academics) feel that principals lack content expertise to accurately evaluate instruction (Nelson, 2010). To the extent this is true, teachers' scores will be less reflective of their true teaching capacity, rendering consequential decisions less accurate and remediation plans less effective. Second, as others have noted (Henry -Barton, 2010), there are many practical challenges in asking principals to be the sole evaluator for their teachers, especially if repeated observations are necessary and state law calls for evaluating every teacher every year, as in some states. Finally, the fact that principals are so often designated as the sole rater poses a conundrum for states, in that it becomes impossible to not certify principals as raters in this situation.

In fact, details on the process of rater certification and monitoring were absent from most states' guidelines. Currently, only five of the 17 states we studied have guidelines or legislation in place that require raters to meet objective certification criteria beyond attendance at a training session, though in interviews, three more indicated plans to implement rater certification tests in the future. Although the specific certification process is rarely described, policies in those

abovementioned five states all reference certification tests that raters must pass to verify the accuracy and reliability of their scores. In other states, policy-makers interviewed for this study often noted the dilemma they face. One stated that "we don't have a pass score right now because quite frankly we don't know what we would do if we set one and people didn't meet it." Another commented that the initial cut-score set by their state admitted 95% of evaluators; this state's system allowed evaluators to take the certification test multiple times if they did not meet the cut-score the first time. In a related study, a district official noted:

> I mean it's nice to have a policy and have a policy with teeth but in the practical world if a principal doesn't certify, what does that mean for the evaluation process of that school for that year? [Will schools have] somebody else come in and do that? And essentially you've kind of - I mean that has all kinds of implications. You've kind of publicly neutered your principal in front of the faculty. That's going to have implications as far as the leadership.

As this evidence suggests, states face problems in using the inherited system of teacher evaluation at the same time as they strive to improve the rigor and power of that system.

How was the instrument chosen? What was known about the validity and reliability of the instrument prior to adoption?

Reliability and validity are properties of scores, not instruments; the extent to which scores are reliable and valid can reasonably be expected to vary across different contexts, including different rater pools, local instructional guidance and accountability systems, and teacher populations (Hill et al., 2012). Nevertheless, we expected that states using third-party observation instruments might have examined information about the instruments' past reliability and validity in their deliberations about which instrument to adopt. To ascertain this, we first asked an open-ended question to elicit information about how the choice of instruments was

made, then asked a more pointed question regarding whether information about reliability and validity of the instrument was known to state officials.

In response to the first question about the broad process through which an instrument was chosen, only one state respondent reported examining information regarding rater reliability. In seven other states which adopted or developed instruments, respondents noted a combination of two key features of the instrument screening process: scans to determine the degree of alignment between the proposed instrument and existing state teaching and learning standards, and scalability: the potential for training and, in some cases, certifying large numbers of raters quickly. Some respondents also commented on using feedback from key stakeholders, including teachers and unions as well as school and district officials.

Ten states responded to the more pointed question about whether evidence of reliability and validity was examined prior to the adoption of an instrument, or during the development of an in-house instrument. Six of those ten states reviewed such evidence, or investigated reliability and validity issues during a pilot of the instrument. This, in combination with responses to the above question, suggests that although these data were collected or produced by states, they were not major factors in the adoption or design of the observation system. In the case of adopting a third-party system, this makes sense: although existing evidence provided by the developers may be useful during adoption, it would not necessarily transfer to the new state contexts.

Of the states designing their own instrument, the majority reported examining reliability and validity during a pilot phase. In some states where districts had the option to develop their own instrument, rather than adopt the system endorsed or developed by the state, districts were required to complete validation studies for their instrument. Most districts, however, would not

have the resources to replicate the quality of the study conducted by vendors or the state, and thus, even in very decentralized states, districts by and large adopted the system that the state had already deemed as producing valid and reliable scores.

Sampling of lessons: How did you arrive at the number of observation required? Was there a study?

Research around the use of classroom observation instruments suggests decisions regarding the sample of lessons–both their number and timing–are quite important (Bell et al, 2012; Hill, Charalambous, & Kraft, 2012). Collecting data from too few lessons means observers run the risk of mischaracterizing a teachers' practice, for instance by observing two days that both happened to be uncharacteristic of the teacher's practice. However, researchers and practitioners have a strong preference for collecting data from no more lessons than necessary, as this data collection is costly for all involved. In practice, most researchers conduct generalizability studies in order to determine the optimal combination of observations and, in some cases, their timing.

By contrast, state policy-makers reported relying on conventional wisdom, political considerations and sheer practical and logistical realities to make decisions regarding the number of observations per teacher per year. As one interviewee stated, "we have tried to build a system that is practical in terms of being able to realistically implement with [as much] quality and fidelity as possible and still be able to give us the amount of time that we want in terms of observation and data collection." In 16 of the 17 states in our sample, the exact number of observations or the minimum number of observations was defined by law without any evidence as to how that number was chosen. One interviewee who was responsible for the implementation of the evaluation system responded to a question about how the number of observations was

decided by saying "That came through legislation." Although four states cited other studies, in particular the Measures of Effective Teaching (MET) project (Bill & Melinda Gates Foundation, 2011), as influential in determining the number of observations, no state interviewed for this paper conducted a study to determine the number of observations used in their evaluation system.

Instead, practical considerations dominated. One state explained the tension between conducting an optimal number of observations and conducting a minimum number that is practical in the field,

> I don't know if we're going to come up with specific items on how many observations you need [in order] to see the observable parts of our rubric. We're debating that internally as a team right now, because I think we are reading the research from the Gates Foundation and others. You basically need six observations to kind of see a good swath of a teacher's practice. But we're also trying to be realistic and feasible about the paradigm shift that this is, and that we've got a lot of rural districts here where the principal is the only evaluator for 35 teachers, six times per teacher.

Another respondent suggested a trade-off between validity and reliability and quality of implementation: "We really want to build an instrument that will be valid and reliable, but will also practically be able to be implemented in our systems with quality."

In five other states, respondents indicated that the number was determined through a negotiation process with teachers' unions. For example, one state arrived at the number of observations as a compromise between the number recommended by the designer of the instrument and what the teachers' union would agree to. Given these limitations however, two states did specifically acknowledge the benefits of more observations, with one respondent indicating that in the case of highly variable scores for a single teacher (described as scores with a lot of "bounce"), raters were encouraged to conduct additional observations. Likewise, several other states require

additional observations of novice teachers and/or those teachers who receive poor evaluations the prior year.

Overall, the information we were provided suggests that the minimum number of required observations is largely determined by attention to logistical and practical concerns, as well as by negotiations with teachers' unions. While several states seemed to be aware of the recommendations set by other bodies of research--specifically the MET study--there were no immediate plans for states to themselves conduct these types of analyses.

Conclusion

Teacher evaluation systems have undergone marked changes in a very short amount of time. The majority of states we studied are currently piloting new systems or in the beginning stages of full implementation. Despite significant federal funding for these efforts, there remain significant resource constraints, most often felt at the district level where the implementation costs are largely born. These constraints will undoubtedly affect the validity and reliability of the scores produced by the new teacher evaluation systems. However, there remain a variety of system design and implementation decisions that states can optimize to increase the reliability and validity of their teacher evaluation scores even within these constraints.

To achieve such optimization, states need look no further than other states' activities, many of which instantiate best practices. Most states, for example, selected an instrument they expect to meet needs within their context, then engaged raters in training around that instrument. In some states, raters must also pass a certification assessment, and will be required to do so (or retrain) periodically. In one state, evaluators will co-rate lessons with state officials, and four states will

audit scores. In four states, teachers may be observed not only by their principals but also by peers. Finally, five states plan to conduct research on the outcomes of their new system.

Although many states have adopted one or two best practices, these seldom occur as a coordinated program of inquiry and action to achieve reliable and valid scores. While a state may engage in auditing scores, for instance, it may miss the gains to reliability and validity that would accrue from periodic rater retraining and recertification, a stiff program of rater monitoring, and the use of multiple raters per teacher. However, nowhere is this more troublesome than in decisions about which and how many lessons to sample, which are either mandated legislatively, result from practical concerns or negotiations between stakeholders, or, at best case, rest on results from the Measures of Effective Teaching study (Kane & Staiger, 2012). Yet MET findings regarding the relationship between number of lessons and reliability of scores are specific to the scoring design and, to some degree, the district and school contexts included in that study; MET results may not apply to districts using different curriculum materials, operating under different testing and accountability structures, and so forth (for a related argument, see Hill et al., 2012). This suggests that states should more actively investigate the number of lessons and lesson sampling designs required to yield high-quality scores.

The lack of a coordinated program may also have a large impact in the area of consequential validity, in other words, how schools, teachers, and children experience the system. Although many state respondents placed consequential validity – typically in the form of stakeholder opinion – high on their list of criteria for policy success, few described a concrete program of research that would study policy effects. We can imagine two phenomena particularly worth tracing from this point forward: the rate of incorrect decisions (e.g., tenuring poor teachers, dismissing excellent teachers) and the overall impact of the new teacher evaluation systems on

teacher recruitment and attrition. As states and districts seek to recruit the best and brightest into teaching, an evaluation system that is perceived to accurately recognize and reward talent may improve recruitment, just as one that is perceived as being arbitrary and unfair may push candidates towards other career options.

Finally, current reforms in teacher evaluation have potential consequences for a wider set of system-level features, such as teacher professional development initiatives and the resources invested in these efforts, principal training and recruitment as instructional leaders, as well as the day-to-day practice of teaching itself. Whether these reforms positively or negatively affect these features depends, in some part, on the reliability and validity of scores, as well as the design of the accountability system as a whole. Getting the measurement of teaching right, in this view, is critical to improving school and student outcomes.

References

Arizona General Assembly. (2012) Arizona revised statute 15-537: Performance of Certified teachers; evaluations systems. Retrieved from http://www.azleg.state.az.us/ars/15/00537.htm

Arizona State Board of Education. (2011). Arizona framework for measuring educator effectiveness. Arizona State Board of education task force on teacher and principal evaluations

Arizona State Department of Education. (2012). State of Arizona ESEA Flexibility Request. Retrieved from http://www2.ed.gov/policy/eseaflex/approved-requests/az.pdf

Balfour, Moody, Weber, Heath & Cowsert. Georgia Senate Bill 386. Retrieved from http://www1.legis.ga.gov/legis/2009_10/pdf/sb386.pdf

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Bill and Melinda Gates Foundation. (2011). Learning about teaching: Initial findings from the measures of effective teaching project. Retrieved from http://www.gatesfoundation.org/college-ready-education/Documents/preliminary-finding-policy-brief.pdf

Colorado State Board of Education. (2011). Reports and recommendations: State council for educators. Retrieved from http://www.cde.state.co.us/EducatorEffectiveness/downloads/Report%20&%20appendices/SCEE_Final_Report.pdf

Delaware General Assembly. (2011) Title 14: Education Delaware administrative code. Retrieved from http://regulations.delaware.gov/AdminCode/title14/100/106A.pdf

Florida General Assembly. (2011) Statute 1012.34: Personnel evaluation procedures and criteria. Retrieved from http://www.flsenate.gov/laws/statutes/2011/1012.34

Heitin, L. (2011). Evaluation System Weighing Down Tennessee Teachers. *Education Week*, 31(8), 1-15.

Henry-Barton, S. N. (2010, January 1). Principals' Perceptions of Teacher Evaluation Practices in an Urban School District. *ProQuest LLC*,

Hill, H., Charalambous, C., Blazar, D., McGinn, D., Kraft, M., Beisiegel, M., & ... Lynch, K. (2012). Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation. *Educational Assessment*, *17*(2/3), 88-106.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Illinois General Assembly. (2010). Public Act 096-0861. Retrieved from http://www.ilga.gov/legislation/publicacts/96/PDF/096-0861.pdf

Johnston, Spence, Foster, Gibbs, Hodge, King, K., . . . Swalm (2011). Senate Bill 10-191. GeneralAssembly of the State of Colorado. Retrieved from http://www.leg.state.co.us/clics/clics2010a/csl.nsf/fsbillcont3/EF2EBB67D47342CF872576A80027B078?open&file=191_enr.pdf

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues & Practice, 21*(1), 31-41.

Kane, M.(2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). New York, NY: Praeger.

Kane, T. J., & Staiger, D. O. (2012).Gathering feedback for teaching:Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/reports.php

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge: Harvard University Press.

Louisiana General Assembly. (2010) House bill no. 1033. Retrieved from http://www.legis.state.la.us/billdata/streamdocument.asp?did=711248

Martinez, J.F., Borko, H., Stecher, B., Luskin, R., & Kloser, M. (2012). Measuring classroom assessment practice using instructional artifacts: a validation study of the QAS notebook. *Educational Assessment*, *17*(2-3), 107-131.

Maryland State Board of Education. (2012) Maryland ESEA flexibility request. Retrieved from http://www2.ed.gov/policy/eseaflex/approved-requests/md.pdf

Maryland State Department of Education. (2012). Maryland Teacher and Principal Evaluation Guidebook. Retrieved from http://www.marylandpublicschools.org/NR/rdonlyres/167F463A-3628-47B7-8720-353C3216AD1A/32101/MarylandTeacherPrincipalReport_041212_.pdf

Massachusetts Department of Elementary & Secondary Education. (2011) Education laws and regulations: Evaluation of educators. Retrieved from http://www.doe.mass.edu/lawsregs/603cmr35.html

National Education Association. (2011). NEA Teacher Evaluation and Accountability Toolkit. Retrieved from http://www.nea.org/assets/docs/2011NEA_Teacher_Eval_Toolkit.pdf

Nelson, B. (2010). How Elementary School Principals with Different Leadership Content Knowledge Profiles Support Teachers' Mathematics Instruction. *New England Mathematics Journal*, *42*43-53.

New York General Assembly. (2011) Senate bill S6732. Retrieved from http://open.nysenate.gov/legislation/bill/S6732-2011

North Carolina State Board of Education. (2011). Annual teacher evaluation requirement policy. Retrieved from http://sbepolicy.dpi.state.nc.us/policies/TCPC022.asp?pri=02&cat=C&pol=022&acr=TCP.

Ohio General Assembly. (2011*).* House bill 153. Retrieved from http://www.legislature.state.oh.us/BillText129/129_HB_153_EN_N.html

Oklahoma General Assembly. (2010) Oklahoma school code. Retrieved from http://webserver1.lsb.state.ok.us/OK_Statutes/CompleteTitles/os70.rtf

Onecle. Arizona Revised Statutes: Title 15 Education - Section 15-203 Powers and duties. Retrieved November 25, 2012 from the Onecle Wiki: http://law.onecle.com/arizona/education/15-203.html

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123-141.

Rhode Island Department of Elementary and Secondary Education. (2009). Educator evaluation system standards. Retrieved from http://www.ride.ri.gov/educatorquality/educatorevaluation/Docs/EdEvalStandards.pdf

Shavelson, R. J. and Webb, N. M. 1991. *Generalizability theory: A primer* Newbury Park, CA: Sage.

Society for Human Resource Management, (2012). Employee Recognition Survey: The impact of recognition on employee engagement and ROI. Retrieved from http://go.globoforce.com/rs/globoforce/images/SHRMWinter2012Report.PDF

Tennessee General Assembly. (2010). Tennessee first to the top act of 2010. Retrieved from http://www.tn.gov/firsttothetop/docs/First%20to%20the%20Top%20Act%20of%202010.pdf

US Department of Education. (2009). Race to the Top Executive Summary. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

US Department of Education. (2012). Summary of Considerations to Strengthen State Requests for ESEA Flexibility. Retrieved from http://www.ed.gov/sites/default/files/considerations-strengthen.pdf

Endnotes

[1] 138 of 285 total points that could be earned in the RTTT application review process were from the "Great Leaders and Teachers" section.

[2] The thirty-three states (plus the District of Columbia) that have been approved for waivers from NCLB include: Arizona, Arkansas, Colorado, Connecticut, Delaware, Florida, Georgia, Indiana, Kansas, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, Oregon, Rhode Island, South Carolina, South Dakota, Tennessee, Utah, Virginia, Washington and Wisconsin. The eleven states (plus the Bureau of Indian Education and Puerto Rico) with outstanding requests for waivers include Alabama, Alaska, California, Hawaii, Idaho, Illinois, Iowa, Maine, New Hampshire, North Dakota, and West Virginia.

[3] AZ, CO, CT, DE, FL, GA, IL, IN, LA, MA, MD, NC, NY, OH, OK, RI, TN

[4] HI, KY, MN, NJ, NM, PA

[5] DE, GA

[6] AZ, CO, CT, FL, IL, IN, LA, MA, MD, NC, NY, OH, OK, RI, TN

Appendix 1:

Appendix 1 *Examples of "Valid" or "Reliable" Text by State*	
State	Text
Arizona	"The governing board [of the school district] shall prescribe specific procedures for the teacher performance evaluation system which shall include at least the following elements: 1. A reliable evaluation instrument including specific criteria for measuring effective teaching performance in each area of the teacher's classroom responsibility…" (A.R.S. 15-537).
Colorado	"...The frequency and duration of the evaluations, which shall be on a regular basis and of such frequency and duration as to ensure the collection of a sufficient amount of data from which reliable conclusions and findings may be drawn." Also, evaluations should use "evaluation rubrics and tools that are deemed fair, transparent, rigorous, and valid" (Senate Bill 10-191).
Connecticut	"On or before July 1, 2012, the State Board of Education shall adopt… guidelines for a model teacher evaluation and support program. Such guidelines shall [provide guidance on]… a validation procedure to audit evaluation ratings of exemplary or below standard by the department, or a third-party entity approved by the department, to validate such exemplary or below standard evaluation ratings. The State Board of Education, following the completion of the teacher evaluation and support pilot program, pursuant to section 52 of this act, and the submission of the study of such pilot program, pursuant to section 53 of this act, shall validate the guidelines adopted under this subsection." Also, "The teacher evaluation and support pilot program described in subdivision (1) of subsection (a) of this section shall… include a validation process for performance evaluations to be conducted by the Department of Education, or the department's designee…" And lastly, "Upon completion of such study, but not later than January 1, 2014, the Neag School of Education at The University of Connecticut shall (1) submit to the State Board of Education such study and any recommendation concerning validation of the teacher evaluation and support program guidelines adopted by the State Board of Education…" (Senate Bill 458).
Florida	Each district's system must "Include a process for monitoring and evaluating the effective and consistent use of the evaluation criteria by employees with evaluation responsibilities" and "Include a process for monitoring and evaluating the effectiveness of the system itself in improving instruction and student learning" (Florida Statute 1012.34).
Illinois	"…the State Board of Education shall, through a process involving collaboration with the Performance Evaluation Advisory Council, develop or contract for the development of and implement all of the following data collection and evaluation assessment and support systems… A process for assessing whether school district evaluation systems developed pursuant to this Act and that consider student growth as a significant factor in the rating of a

	teacher's and principal's performance are valid and reliable, contribute to the development of staff, and improve student achievement outcomes. By no later than September 1, 2014, a research-based study shall be issued assessing such systems for validity and reliability" (PA 096-0861).
Louisiana	LEAs must provide "an explanation of how the LEA will ensure the reliability and validity" of an alternative evaluation system they would like to use (Bulletin 130).
Maryland	There is no "valid" or "reliable" text in Maryland's Education Reform Act or Title 13A. But the executive order which established Maryland's Educator Effectiveness Council states that "The Council's recommendations should seek to ensure that every educator is... Evaluated using multiple, fair, transparent, timely, rigorous, and valid methods..." (Educator Effectiveness Council Executive Order).
North Carolina	"A local board of education shall use the North Carolina Professional Teaching Standards and North Carolina Teacher Evaluation Process unless it develops an alternative evaluation that is properly validated and that includes standards and criteria similar to those in the North Carolina Professional Teaching Standards and North Carolina Teacher Evaluation Process" (Teacher Evaluation Process).
Oklahoma	"There is hereby created to continue until July 1, 2016, in accordance with the provisions of the Oklahoma Sunset Law, the Teacher and Leader Effectiveness Commission… The Commission shall provide oversight and advise the State Board of Education on the development and implementation of the Oklahoma Teacher and Leader Effectiveness Evaluation System (TLE) as created in Section 6-101.16 of this title, including… Regularly reviewing progress toward development and implementation of the quantitative and qualitative measures that comprise the TLE;… Regularly reviewing the correlation between the quantitative and qualitative scores and other data to ensure that the TLE is being implemented with validity and that evaluations of individuals conducted by school districts are meaningful and demonstrate that reasonable distinctions are being made relating to performance;… Assuring that the TLE is based on research-based national best practices and methodology… The Commission shall issue a report by December 31 of each year and submit a copy of the report to the Governor, the Speaker of the House of Representatives and the President Pro Tempore of the Senate" (Oklahoma OS70).
Rhode Island	"The evaluation system provides safeguards against possible sources of bias to ensure valid assessments. Districts review evaluation instruments for possible sources of bias in the design process and monitor implementation results for possible inappropriate adverse impact. Evaluators raise existing or potential conflicts of interest so they can be addressed. The evaluation system provides procedural safeguards (e.g., appeals) to ensure the integrity of the system" (Rhode Island Educator Evaluation Standards). Also, "Districts establish and support a District Evaluation Committee that includes teachers, support professionals, administrators, and union representatives… The Committee reviews the effectiveness of the evaluation system, the validity and utility of the data produced by the system, the fairness, accuracy, and consistency of

	decisions made, and the currency of the system. The Committee uses the information from the analysis to make recommendations for revisions to the system" (Rhode Island Educator Evaluation Standards).

Exhibit 1 *Characteristics of Evaluation Systems by State*		
Question answered by the data	States	Notes
What is the breakdown of local versus state control in the teacher evaluation system? (opinion)	**Entirely state-based:** DE, GA **State model or choice of models with some district choice:** IL, MD, MA, NC, OH, OK, RI, TN **Mainly district developed:** AZ, CO, CT, FL, IN, LA, NY	
Which states specify consequences for poor performance?	All states besides for Arizona (16 out of 17) specify consequences of poor performance on a teacher evaluation. Consequences range from not getting raises, assistance programs, and remediation plans to dismissal or termination.	
Which states specify consequences for excellent performance?	**None:** CT **Some:** AZ, CO, DE, FL, GA, IL, IN, LA, MD, MA, NY, NC, OH, OK, RI, TN	Consequences of excellent performance range from nothing specified by the state to bonuses, salary increases, non-probationary status, tenure, and fewer observations in the future.
Does the state have legislation or regulation using the terms "reliable" and "valid"?	**Yes:** AZ, CO, CT, IL, LA, MD, NC, NY, OK, RI **No:** DE, FL, GA, IN, MA, OH, TN	This binary binning does not including mentions of "valid" or "reliable" in NCLB waivers or Race to the Top applications.
Does the state specify that there should be inquiry into the reliability and validity of scores?	**Yes:** CO, CT, GA, IL, LA, MA, NY, NC, OH, OK, RI, TN **No:** AZ, DE, FL, IN, MD	

Does the state specify that there should be inquiry into the reliability and validity of the VAM scores	**Yes:** CO, GA, MA, NY, TN **No:** AZ, CT, DE, FL, IL, IN, LA, MD, NC, OH, OK, RI	The state must mention that they are looking into the validity and reliability of the VAM or the 'components' of the teacher evaluation system.
Does the state specify that there should be inquiry into the reliability and validity of the observation system?	**Yes:** CO, GA, LA, MA, NY, OH, TN **No:** AZ, CT, DE, FL, IL, IN, MD, NC, OK, RI	The state must mention that they are looking into the validity and reliability of the observations or the 'components' of the teacher evaluation system.
Does the state have legislation requiring a report on aspects of reliability or validity?	**Yes:** CT, IL, OK **No:** AZ, CO, DE, FL, GA, IN, LA, MD, NY, NC, OH, RI, TN	
Who observes the teacher? (Binned)	**No specification:** AZ, CT, GA, OK **Principal, administrator, supervisor, instructional leader:** DE, FL, LA, MD, TN **Superintendent, principal, administrator, or designee:** CO, IN, MA, NY **Other:** IL, NC, OH, RI	
Does the state require multiple raters?	**Yes:** NC **No:** AZ, CO, CT, DE, FL, GA, IL, IN, LA, MD, MA, NY	NC - Multiple raters are only required for probationary teachers, who receive one peer observation in addition to three evaluations from a principal. For career status teachers, the peer observation is dropped. IN - Although Indiana does not require multiple raters, they do say "Corporations may want to consider: Allowing for second or third party observers to provide multiple perspectives. In collecting evidence of teaching practice, it is not only important to use multiple sources of evidence or multiple measures, it can also be helpful to both evaluator and teacher if a second or third

		party observes" (Legislation Guidelines Evaluation Plans). Maryland recommends additional raters in the case that a teacher is scores in the ineffectual category
Who does the state specify should train evaluators?	**State training:** DE, GA, IL, MD, MA, OH, TN **District training:** CO, CT, FL, IN, LA, NY **No specification (or other):** AZ, NC, OK, RI	Maryland trains executive officers who in turn train principals to evaluate teachers.
Does the state have certification?	**Yes:** DE, GA, IL, LA, MD, NY, NC, OH, OK, TN **No—but there is training:** AZ, CO, CT, FL, IN, MA, RI	Delaware's evaluators are "trained and credentialed" (DPASII Full Guide). IL - Any evaluator undertaking an evaluation after September 1, 2012 must first successfully complete a pre-qualification program provided or approved by the State Board of Education. The program must involve rigorous training and an independent observer's determination that the evaluator's ratings properly align to the requirements established by the State Board pursuant to this Article" (PA 096-0861).
Does the state have legislation or guidelines which require raters to meet objective certification or training criteria other than simply attending a training workshop?	**Yes:** IL, LA, OH, TN, OK **No:** AZ, CO, DE, GA, CT, FL, IN, MA, MD, NC, NY, RI	
Among the 17 examined states, what number of observations per teacher per year of either a formal or informal nature is specified?	**None:** OK **1:** FL, DE **2:** AZ, CO, GA, IL, IN, LA, MD, NY, OH **3:** CT, MA, NC, RI	In Oklahoma, no number of observations is specified by law or guidelines, but LEAs must choose from the Marzano, Danielson, or Tulsa teacher evaluations systems, each of which has its own guidelines. Florida requires one or two observations depending on whether the teacher is a novice or not.

	4: TN	Delaware requires one to three observations depending on the experience and previous evaluations of teachers. Illinois requires two or three observations depending on the experience and previous evaluations of teachers. Ohio also requires classroom walkthroughs. Massachusetts requires three or five observations depending on the experience and previous evaluations of teachers. North Carolina requires three observations of career status teachers and four observations of probationary teachers (one of which is a peer observation).
Are all teachers evaluated once or more per year?	**Yes:** AZ, CO, CT, FL, GA, IN, LA, MD, NY, NC, OK, RI, TN **No:** DE, IL, MA, OH	MD - Certain teachers may be evaluated using part of their evaluation from the previous year. Tenured teachers rated as highly effective or effective must use the most recently calculated student growth measures each year, but they may reuse the professional practice rating from the previous year instead of having to be reevaluated on the professional practice dimensions (Maryland Teacher Evaluation Guidebook). DE - Teachers may receive a one year waiver. IL - Tenured teachers who achieve one of the highest two possible ratings may be evaluated only once every two years. MA - Educators will receive a summative evaluation every one to two years depending on experience and previous evaluations. OH - The board may elect, by adoption of a resolution, to evaluate each teacher who received a rating of accomplished on the teacher's most recent evaluation once every

		two school years.