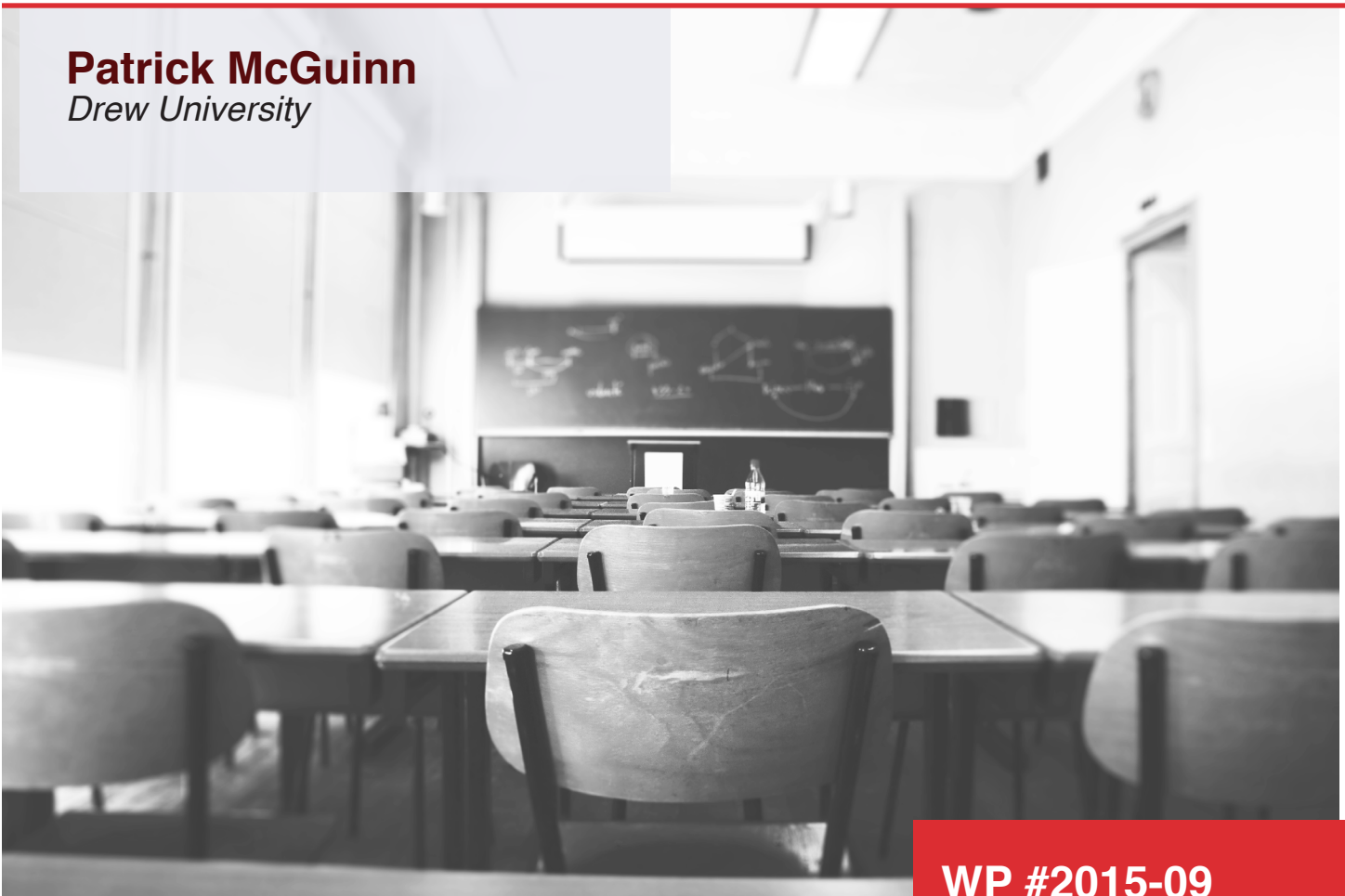


SEPTEMBER 2015

Evaluating Progress:

State Education Agencies and the Implementation of New Teacher Evaluation Systems

Patrick McGuinn
Drew University



WP #2015-09

Suggested Citation:

McGuinn, P. (2015). *Evaluating Progress: State Education Agencies and the Implementation of New Teacher Evaluation Systems*. White Paper (#WP2015-09). Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania.

About the Author

Patrick McGuinn is Associate Professor of Political Science and Education and Chair of the Political Science Department at Drew University and a Senior Research Specialist at the Consortium for Policy Research in Education (CPRE). He holds a Ph.D. in Government and a M.Ed. in Education Policy from the University of Virginia. Patrick previously held fellowships at the Institute for Advanced Study in Princeton, the Taubman Center for Public Policy at Brown University, and the Miller Center for Public Affairs at the University of Virginia, and was a visiting scholar in the Education and Politics program at Teachers College, Columbia University. His first book, *No Child Left Behind and the Transformation of Federal Education Policy, 1965-2005*, (Kansas, 2006) was honored as a Choice outstanding academic title. He is also the editor (with Paul Manna) of *Education Governance for the 21st Century: Overcoming the Structural Barriers to School Reform* (Brookings Institution Press, 2013). He is currently working on a new book project with Chris Loss of Vanderbilt entitled *Convergence: U.S. K-12 and Higher Education Policy Fifty Years after the ESEA and the HEA of 1965* which is under contract at Harvard Education Press. Patrick has published many academic articles and book chapters and has produced a number of policy reports for the Brookings Institution, the American Enterprise Institute, the Center for American Progress, and the Thomas B. Fordham Institute. He is a regular commentator on education policy and politics in media outlets such as *Education Week*, the *New York Times*, the *Wall Street Journal*, and the *NJ Star Ledger* and was named as one of the top “Edu-Scholars” in the country by *Education Week* for the past four years. He is a former high school social studies teacher and the father of four daughters who attend public schools.

About the Consortium for Policy Research in Education

The Consortium for Policy Research in Education (CPRE) brings together education experts from renowned research institutions to contribute new knowledge that informs PK-16 education policy and practice. Our work is peer-reviewed and open-access at cpre.org. CPRE’s member institutions are the University of Pennsylvania; Teachers College, Columbia University; Harvard University; Stanford University; University of Michigan; University of Wisconsin-Madison; and Northwestern University.

Introduction

In a 2012 paper for the Center for American Progress, “The State of Evaluation Reform,” Patrick McGuinn (Drew University) identified the opportunities and challenges facing education agencies in Race to the Top (RTTT) grant-winning states as they prepared for the implementation of new teacher evaluation systems. Three years have now passed, and states have moved from planning and piloting to full implementation of the new systems. Yet a survey of media coverage in 2014-2015 reveals that while many states have made considerable progress in rolling out their new evaluation systems, struggles remain and most grantees have asked to extend the timetables for completing this work.¹ An April 2015 GAO report concluded that RTTT grantees “noted various challenges to their capacity to successfully support, oversee, and implement these reform efforts.”² Given the enormous importance and complexity of these reforms—and the fact that states vary widely in the timing, approach, and success of their implementation work—this is an excellent opportunity to assess the progress that has been made and identify where challenges persist. It is imperative that states learn from one another during this implementation stage, and this report hopes to facilitate that by highlighting what is and is not working in the field.

The 2012 study undertook in-depth comparative case studies of six states: Tennessee, Colorado, Delaware, New Jersey, Rhode Island, and Pennsylvania. These particular states were selected because they were “early adopters” in the area of teacher evaluation reform and because their states and/or education agencies had undertaken different approaches to implementing the reforms. For this paper the individuals interviewed in those states two years ago (or their replacements if necessary) were re-interviewed to understand how and why their efforts differ today. By analyzing state implementation efforts at two different points in time, the new study utilizes a longitudinal qualitative approach that can reveal the extent to which states are learning and adapting in this work over time. Rather than the detailed state case studies of State Education Agency (SEA) implementation work provided in the 2012 paper, this report uses a more thematic approach that will synthesize the lessons that have emerged from the field. Research consisted of a review of the scholarly and think tank research on SEA capacity and teacher evaluation systems; analysis of reports and data from the state education departments’ websites, the U.S. Department of Education, and from organizations such as the Council of Chief State School Officers (CCSSO) and National Council of Teacher Quality (NCTQ); media coverage of the reform efforts in the case study states; and, interviews with SEA and local education agency (LEA) staff in the original case study states.

It was clear that SEAs were working hard in 2012 to realign their organizations with the many new

FREQUENTLY USED ACRONYMS

In this paper, we have used each state’s unique acronym to refer to its department of education:

Colorado: CDE

Delaware: DDOE

New Jersey: NJDOE

Pennsylvania: PDE

Rhode Island: RIDE

Tennessee: TDOE

¹ See Michelle McNeil, “Race to Top Reports Detail Winners’ Progress, Challenges: Teacher Evaluation Puzzle Proving Difficult to Crack,” *Education Week*, March 27, 2015.

² Government Accountability Office, “Race to the Top: Education Could Better Grantees and Help Them Address Capacity Challenges,” April 2015. <http://www.gao.gov/products/GAO-15-295>

responsibilities that had been thrust upon them in the wake of No Child Left Behind (NCLB) and RTTT. State efforts to implement new teacher evaluation reforms offer an excellent case study of the ways that SEAs were adapting to their new role as well as the ways in which ongoing capacity gaps continue to impede their work. Improving teacher quality has become the centerpiece of the Obama education agenda and of the contemporary school reform movement. The many challenges that have already emerged, however, also highlight how difficult this work is and how it is complicated by short timelines and limited SEA staff and funding.

The purpose of this paper is twofold:

1. To provide a snapshot in time (Jan 2015) of SEA implementation efforts around new teacher evaluation systems.
2. To contrast more recent implementation efforts with those two years earlier to understand the ways in which SEAs have (and have not) learned and adapted their implementation work over time.

More specifically, the paper will address the following questions: What kinds of capacity—financial, personnel, technical—have SEAs added to support the implementation of new teacher evaluation systems? What kind of capacity is still lacking? How rapidly and how effectively are states implementing their new teacher evaluation systems? Why do some states appear to be having more success/smooth implementation than others? How are states approaching this implementation work differently from one another—do some approaches appear to be more or less effective than others? What challenges are emerging and how are states addressing these? What lessons can be learned from these “early adopter” states that can inform teacher evaluation reform in the rest of the country? How are states approaching the training of evaluators and the principals and teachers who are supposed to use the evaluations to improve personnel decisions and classroom instruction? How well are new teacher evaluation systems being aligned with other reforms such as the move to Common Core and new assessments? How are states dealing with the challenge of measuring student achievement in non-tested subjects?

In 2012, McGuinn’s “The State of Evaluation Reform” paper identified the following key challenges in implementing new teacher evaluation systems:

The Philosophical/Statutory/Constitutional Debate over the Proper SEA Role

Each state’s education agency has a unique history and operates in a different fiscal, political, statutory, and constitutional context. In particular, states vary significantly in their attachment to local control of schools and the proper role of the state in education and this had a major impact on how SEAs were approaching teacher evaluation reform. SEAs’ traditional focus on compliance and accountability activities was making district officials wary of being candid about whether and how they might be struggling to implement reform and reluctant to seek out assistance.

The Amount of Flexibility in State Evaluation Systems Varies Greatly

States vary widely in the amount of centralization/standardization they have mandated—either in statute or in regulation—in the new teacher evaluation systems, and this was having a major impact on the SEA’s approach to supporting implementation. A clear tension was apparent between states’ desire to give districts flexibility to select or adapt evaluation instruments that are best suited to their particular circumstances, and SEAs’ limited capacity to provide implementation support for a wide array of instruments.

SEA Restructuring and the Human Capital Demands

In 2012, SEAs in many states were undergoing a radical restructuring and re-staffing as they embraced a shift from being compliance monitors to being service delivery/school improvement organizations. This restructuring was difficult and time-consuming work and while necessary to carry out new responsibilities over the long term, in the short term it was creating a number of challenges. Many SEAs had created new Teacher Effectiveness units, but the degree to which these units were being well integrated with other units varied and longstanding concerns about agency siloing persisted.

Internal vs. External Capacity

During the early stages of planning and piloting in 2012, SEAs dealt with their internal capacity gaps by relying on two different kinds of external capacity: outside consultants and foundations. There was some concern, however, that reliance on outside grants and consultants would preclude or delay the development of the fiscal self-sufficiency and internal capacity that could support these systems over the long term.

Funding Streams and the Fiscal Cliff

There was a great deal of concern about SEA's lack of capacity to implement these reforms, particularly for states that did not win a RTTT grant or secure foundation support (which is the majority of states). Given the tight fiscal climate in 2012, most states were unable to allocate new money to support the implementation of these reforms. SEAs appeared to vary widely in the way that they had spent external funds, the degree to which they were dependent on them, and the extent to which they had begun to bring these expenses on budget.

Evaluating the Evaluators

One of the primary activities of SEAs in preparing school districts for the implementation of the new teacher evaluation systems was to provide training to the administrators who would be conducting the new observations. States varied widely in their approach here, however, for both philosophical and capacity reasons with some SEAs (such as Tennessee) directly training all evaluators, some (such as Colorado and Pennsylvania) adopting a train-the-trainer model, and others (such as New Jersey) leaving the training entirely up to districts.

Implementation Timetables and Sequencing

Most state reform statutes established rapid timetables for the installation of new teacher evaluation systems. While all states were struggling to meet these timetables, it was clear that some states were struggling more than others and that this was related to the fact that states vary in terms of their experience with statewide evaluation systems. A related challenge centered around the extent to which evaluation reforms were—or were not—being connected to the implementation of other reforms such as new principal evaluations, and new Common Core State Standards and assessments.

Value-add/Growth Scores for Teachers in Non-Tested Subjects

Perhaps the single biggest challenge in implementing new evaluation systems that emerged in 2012 was the fact that the majority of teachers did not teach in tested subjects or grades and thus standardized student achievement data was not available to be used in their ratings. Districts were working independently to develop their own student learning objectives (SLOs)—but the quality of the results appeared to be very mixed and messy both within and across states. This was an enormous problem and it was clear that many SEAs were struggling to address it.

Networks, Policy Learning, and Politics

Policy learning and continuous improvement requires that districts, SEAs, and the U.S. Education Department be transparent and forthcoming about what is working and what is not and that lessons learned be regularly shared within and between states. However, the reality in the field was that not enough communication and sharing of information about what worked and did not work was occurring in 2012.

Issues that Emerged from 2014-15 Interviews and Research

ONGOING SEA AND LEA CAPACITY ISSUES

In 2012, building the capacity at both the district and SEA level to implement new evaluation systems was a major issue across the country. States and districts struggled to secure the financial, personnel, and technical resources to support implementation. The economic downturn and budget cuts had led to staff cuts in many places at exactly the moment when additional personnel were needed to carry out the demanding new evaluation work. The staff/capacity issue continues to be exacerbated by the way many SEAs and districts are structured around discrete funding streams which leads to a serious ‘siloeing’ problem and makes it difficult to re-assign staff to new functions.³ Despite the clear need for SEAs and districts to provide sustained support to schools, the significant capacity issues that were identified in 2012 remain today.⁴ Daniel Weisberg from the New Teacher Project (TNTP) believes “that capacity is a huge challenge at the state level —state departments of education often just don’t have the resources to really do a full state-wide rollout of a major initiative and ensure quality implementation in every district. Race to the Top required them to go beyond policy to actually be the implementers and that’s a very different role.” Michelle Exstrom of the National Conference of State Legislatures (NCSL) concurs, noting that:

There’s still quite a bit of concern that the SEAs are doing this on a shoestring budget with probably the fewest amount of staff that they can possibly spare to be focusing on these things, and that’s probably not an ideal situation in most states. Even if the budgets are in great shape, we don’t see them providing significant appropriations to SEAs for things like this so they’re still having to figure out and piece it together.

In Pennsylvania, David Volkman, Special Assistant to the Acting Secretary in the PA Department of Education (PDE) noted that:

Our agency (PDE) has shrunk by over 50% in just the last six years and by that I mean in terms

3 Ashley Jochim and Patrick Murphy, “The Capacity Challenge: What it Takes for State Education Agencies to Support School Improvement,” Center on Reinventing Public Education, December 2013.

4 Center on Education Policy, “State Education Agency Funding and Staffing in the Education Reform Era,” February 2012.

of personnel that we have on board. We really do lack capacity in terms of the number of staff members who can effectively manage many of these very, very important projects. So, in Pennsylvania, what we have come to do is to rely heavily on our Intermediate Units—we have 29 of them—and then we also bring contractors to the table.

Money was also put into this project in PA from Teacher Professional Development dollars, a line item that they get from the legislature as well as from their RTTT grant and some other federal dollars. But Volkman (PDE) added that:

Honestly we have limited capacity and I think you'd find the same in any other state and I think that's critical sometimes because when you look to hand the project off after it's developed and it's implemented, who's going to do the monitoring? Who's going to do the maintenance, if you will? Who's going to go back and revisit that and make those necessary modifications moving forward?

Some SEAs,—particularly those with RTTT or foundation grants—have managed to add or re-assign staff to manage the teacher evaluation work. Other SEAs have partnered with outside organizations to supplement their in-house capacity, as Tennessee has done with TN SCORE.

The 2012 CAP report noted that many SEAs were restructuring themselves and creating new Talent or Human Capital offices to support districts with the roll-out of new evaluation systems and this work continues.⁵ Janice Poda from the CCSSO remarked in 2014 that:

Almost every SEA now has personnel designated to help districts with this work. Before only a few states had personnel at the SEA designated to assist districts with teacher evaluations especially in places that have local control and traditionally had left teacher evaluations up to the districts as a local responsibility. And there was no one that was really spearheading teacher evaluation at the state level.

Another positive development is that the high turnover among the SEA staff doing the educator evaluation work observed in 2012 seems less prevalent in 2014. In the six case study states most of the folks leading this effort remained in place from two years ago. In New Jersey, Rhode Island, Colorado, and Tennessee, the SEA leadership on evaluation remains the same while in Delaware and Pennsylvania there was turnover.

It appears that a similar kind of restructuring has been occurring within some district central offices over the past two years as they have gotten farther along in the implementation work.⁶ In Metro Nashville Public Schools (MNPS), for example, in January 2013 the Department's old Human Resources office was renamed "Human Capital" and re-focused on developing a talent strategy for the state and on recruiting, supporting, and retaining great teachers. Shannon Black, the district's Director of Talent Management, notes that "My responsibilities largely entail the performance management piece, really becoming the point person and the driver of the evaluation work in the district. I support our central office staff, our leadership and learning division around teacher evaluations as well as

5 For more ideas on how SEAs should restructure themselves to meet their new responsibilities see: Andy Smarick and Juliet Squire, "The State Education Agency: At the Helm Not the Oar," Fordham Institute, April 2014. <http://edexcellence.net/publications/the-state-education-agency-at-the-helm-not-the-oar>

6 Karen Shakman, Nicole Breslow, Julie Kochanek, Julie Riordan, and Tom Haferd, "Changing Cultures and Building Capacity: An Exploration of District Strategies for Implementation of Teacher Evaluation Strategies," Education Development Center, 2012 Accessed online on August 1, 2014 at <http://tld.edc.org/resource-library/district-strategies-implementation-teacher-evaluation-systems>

ongoing communications with principals and teachers. It is important for me to gain perspectives on the ground around teacher evaluation.” While some large urban SEAs like Nashville may have the resources to create such human capital and talent management offices, it seems unlikely that small school districts—which constitute the majority of districts in the country—will have the capacity to do so. This emphasizes the need for SEAs to differentiate their support to meet the distinct needs of different kinds of districts. Even in large urban districts such as Nashville, however, the Human Capital office is often just a “team of one” as Black (MNPS) noted.

It does not appear that most districts are adding much if any capacity or staff, but instead relying more on principals and other school building administrators to do this work. This in turn, however, has led to concerns that principals are being overloaded and unable to devote the necessary time to doing the evaluation work. Tim Gaddis, Teaching Learning and Assessment Director from the Williamson County district in Tennessee acknowledged that:

They’re overburdened, that’s for sure. You know, our principals, they simply cannot get all the work done in a reasonable work week, and so they’re working far beyond what is reasonable. I think our principals have really bought into the notion that, look, this is hard, but it’s worth it; we’ve seen great achievement gains in our district the past two years. But it’s an overwhelming job. What ends up happening is they spend a lot of time at night in their buildings doing those things that maybe at one time could’ve come during the day.

This dynamic may well prove unsustainable, however, as existing principals may prove prone to burnout and prospective principals are turned off by the prospect of longer hours and inadequately supported work.

Limited SEA resources combined with widely divergent district needs around implementation support have led many state agencies to differentiate and prioritize the kinds of support they provide. New Jersey’s Peter Shulman, for example, commented that:

We have close to 600 school districts, and they have a diversity of needs and diversity of challenges. And when we think about the support, we think about the ability to actually be hands-on with districts, we want to make sure that the support is wherever possible tailored to the individual needs of the district. So if you think about different demographics, different socioeconomic problems, different sizes, we’ve really tried to make sure that as we deploy our resources, we do so with that lens in mind.

Similarly, Weisberg (TNTP) believes that “Rather than using their limited resources to provide relatively light-touch support to all districts, it may be more effective to differentiate support and to provide significant support to a few districts in order to create exemplar districts. It is important to create some real success stories and some proof points that other districts can look to in order to see what’s possible.”

Like Tennessee, New Jersey undertook (in 2011) a major restructuring of its department of education that led to the creation of a new Human Capital office, along with a variety of new units to provide support to school districts. Peter Shulman (Chief Talent Officer/Assistant Commissioner of Teacher and Leader Effectiveness for the NJ Department of Education [NJDOE]) stated that “we wanted to think about how we sort of deploy our resources in a disproportionate manner; it really concentrates on the folks that need them the most.” He emphasized that they think about implementation in terms of four tiers of state support: the state agency, implementation managers, the county offices, the regional achievement centers (RACs).

So by having these multiple tiers for support, I'm a district out there and I have a concern. I have the ability to first go to my county office. Or if I have a RAC working with me I'm going to go to the executive director of the RAC. If it's a simple question that can be answered through a regulatory response, and simply out there as a black and white issue, they'll handle it on the ground, right away. If it's more technical and saying, hey, we're having trouble with SGOs or a schedule for administrators to conduct the required number of observations, we deploy implementation managers. We have three of them across the state, one in the north, one in the center, and one in the south, who literally go to districts, go to schools, go to classrooms, to meet individual needs where our RACs and county officers aren't able to.

New Jersey also created new structures inside of every school and district in the state by requiring (in statute) a School Improvement Panel and (through regulations) a District Evaluation Advisory Committee. Tim Matheny, the former Director of Evaluation remarked that:

One thing that we found in the course of the first year of implementation is that school districts that thoughtfully, deliberately, and collaboratively made policy decisions around their district evaluation systems had much greater success, had much less conflict around the evaluation process. So districts that have really collaborative processes around these—these two groups, we have found really good implementation at a very high level. So we really believe that these two groups, the School Improvement Panels and the District Evaluation Advisory Committees have a lot of potential for making evaluation collaborative and thoughtful and deliberative, which is what we certainly want at the state level.

NJ's use of multiple levels of support and multiple structures for stakeholder engagement and deliberation seems to offer a promising approach to collecting and disseminating information and promoting effective implementation.

BRINGING GRANT-FUNDED PROGRAMS ON BUDGET

Many SEAs and districts have relied on either federal or foundation grants to build capacity and support the early implementation of new teacher evaluation systems. In Tennessee, for example, the Teacher Evaluation unit at the state department of education had only one employee prior to RTTT but they were able to expand that into a team of four in the central office supported by 18 coaches that work regionally. They also used those funds to build a data system that districts can use so that they do not have to have their own data system for collecting observation scores and running reports. But as in many states, that external funding is now coming to an end. Thus it is imperative for states and districts to develop a sustainability plan that identifies what kinds of support should be continued over the long-term and how that support will be funded. In Tennessee, the Department put forward a budget request to the TN General Assembly (which was approved) for \$1.3 million to continue much of the evaluation work using state dollars. Tennessee's example aside however, additional dollars seem unlikely to be forthcoming to support implementation in most states due to tight budgets.

Mary Ann Snider, the Chief of Educator Excellence and Instructional Effectiveness for the Rhode Island Department of Education (RIDE) noted, for example, that:

We were able to add capacity through our race-to-the-top grant, but as the grant concludes we

face challenges to maintain the level of support necessary to continue high-quality work with our educators and to maintain our technology systems. The types of capacity that we think are absolutely essential include data analyses and technology expertise. We believe that our role is to support districts' use of data in order to inform their decisions about educator placement, retention, and professional development. And we need people with evaluation content expertise to help continue to refine the model, lead advisor group meetings, and provide professional development to the field.

She reported that RI has a sustainability plan and that they're going to be able to keep a couple of core people who were originally funded by the state's RTTT grant, but not as many as they would like or need. But it is also clear that the role of the SEA in educator evaluation may change after the new systems are fully developed--the start-up phase of this work is likely to differ significantly from the maintenance phase. Snider (RIDE) remarked that:

I believe going forward we are asking ourselves, with reduced state involvement, what is an appropriate SEA role? Is our role to make sure that districts fully own the system with our taking on a monitoring role, or is there some need for us to provide ongoing professional development to ensure that the system is continuously improving? So, we're working with the districts to sort that out and plan appropriately.

States, districts, and schools are going to have to free up resources and personnel by reorganizing or finding new efficiencies. A recent study found that states could help districts more efficiently and effectively utilize existing dollars by eliminating (or waiving) time-consuming regulations and reporting requirements and giving them greater flexibility with how they allocate their budgets.⁷ The Colorado Education Initiative (CEI), for example, has done a study of four small and large districts in Colorado to investigate in depth where they are spending their money and what their return on investment is for those allocations. Mike Gradoz, Director of Educator Effectiveness at CEI, noted that "what we're trying to do is help districts think differently about reallocating resources and prioritizing what they want to sustain that they believe is really critical to them in terms of implementing the evaluation and standards and assessments." A recent GAO survey of SEA officials found them very worried that fewer resources and staff once RTTT grants expired would have a detrimental effect on future implementation efforts.⁸ Making matters worse is that the federal contract for the Reform Support Network—which had been providing technical assistance to RTTT states—expired in September 2014. Federal policymakers should consider allowing states and districts to re-allocate funds from other federal programs to support this work over the longer term.

EVALUATOR TRAINING AND CERTIFICATION

It has become clear that the training of evaluators is the single most important task in the implementation of new teacher evaluation systems. But there are ongoing concerns about the quality

7 Karen Hawley Miles, "Promoting Transformation Through the State Budget Process and Policy," PIE Network, 2014.

8 Government Accountability Office, "Race to the Top: States Implementing Teacher and Principal Evaluation Systems Despite Challenges," September 2013.

and variability of evaluator training within and across states.⁹ States are approaching this task in very different ways, with some states like RI and TN having the SEA train all evaluators directly, while other states are utilizing a train-the-trainer model, and some states like New Jersey are leaving this training entirely up to districts. In NJ, as Tim Matheny (Director of Evaluation, for NJDOE) notes,

We do not train directly on practice instruments because we have twenty-five different models, and districts get to choose which ones that they adopt. Districts have the responsibility to train their observers and teachers on those practice instruments. And ultimately superintendents make a decision about who is qualified in their district to conduct observation obviously within the parameters of regulation and statute.

As Poda (CCSSO) commented, “it really goes back to the state’s philosophy on whether it’s more centrally controlled or if it’s left to the local decision makers. Anywhere you have greater local authority and traditional local control, you’re going to see less and less participation in those kinds of training.

The RIDE has developed a promising approach to ongoing training around the evaluation work. Every summer they run training institutes for all evaluators: a two-day session for veteran principals and a four-day session for people who are new to being a principal or to an evaluation role and need more comprehensive training. This past year the Department also offered “calibration sessions” during the academic year where they had a team from their office go into a district and work with their leadership team. So everyone who was part of an evaluation system went through the same calibration session, and they offered four of them; districts could choose to participate in a minimum of two, but could do all four if they wanted. The sessions focused on setting student learning objectives, observing teachers, providing feedback, and scoring learning objectives. For the scoring session, they ask that districts send the Department in advance a representative sample of student learning objectives that teachers had created and that they approved. Snider (RIDE) notes that:

Last year we dedicated significant time to hosting in-district calibration sessions on observations and SLO scoring. We used videos from multiple sources, not just those used in our online training tool to be used during the sessions. We had the principals and central office leaders score them in a group, share their ratings, and talk about how they scored them and why. Those were really, really productive. People really appreciated being part of a training that was in their district, just for their team. It revealed discrepancies in judgment among a district evaluation team and reinforced where they were aligned. I think it felt like a safer environment for them to have those conversations and in the end made them much more confident and reliable in their ability to observe practice.

In PA, evaluators train on Teachscape and they get a score at the end to determine where they land but that’s only for their own information. Many districts in PA are having their principals go through the evaluation training together as a cohort. Patricia Hardy (Teacher Effectiveness Project Lead Consultant in the PDE) observed that “that this led to collaboration and conversations around ‘well, what did you see’ because they’re looking for evidence. This is really different and a big shift for some of these people—looking for evidence that’s related to a specific component to Danielson [PA Teacher Observation Framework].” The Department also partnered with the Pennsylvania

9 Grover Whitehurst, Matthew Chingos, and Katharine Lindquist, “Getting Classroom Observations Right,” *Education Next*, Winter 2015, Volume 15, No. 1. <http://educationnext.org/getting-classroom-observations-right/>

Association of Elementary and Secondary School Principals and is developing additional training tools for which people can get 45 professional development credits. The PDE are also going to be giving the principals association their own cadre of trainers to use. PDE regularly surveys their districts and intermediate units and conducts focus groups to better understand what is going on in the field. They intend to use the aggregate data on principals' evaluations (which will be available for the first time this year) to develop trend data and identify where additional professional development may be necessary.

The PDE also works with the state's teacher union, the Pennsylvania State Educators Association and other professional organizations to identify teachers who are struggling with the Danielson Framework or figuring out how to adapt it to their particular work. These teachers were then recruited to volunteer for additional training. Angela Kirby (Director of the Harrisburg Training and Technical Assistance Network) noted that:

We brought those educators together, we had them review the instruments, we had them develop examples of how the instruments could apply to their specific roles and functions across the levels of proficiency, we had them work on the development of guiding questions that could support their supervising administrators and eliciting conversations around their roles in practice. Additionally we also worked with specialist workgroups, folks who were in non-teaching roles but also critical related service providers in our schools—our school counselors, our school psychologists, our school nurses. For those individuals, we worked collaboratively on the development of their own specific evaluation instruments. So it's really been a process that has engaged stakeholders, practitioners in the field to do this work so they see themselves represented in this work.

The importance of giving educators hands-on training and the opportunity to provide feedback on the training and evaluation instruments was a common refrain in the interviews for this paper.

The Colorado Department of Education (CDE) partnered with My Learning Plan to create "Elevate Colorado," an online inter-rater agreement training system "to promote common interpretations of teacher quality and help evaluators provide useful and actionable feedback to educators." Using the system, educators can view short videos of practicing teachers, align observable professional practices from the state rubric and then receive feedback showing how close their alignment is compared to that of master scorers.¹⁰ The Colorado Education Initiative (CEI) also created a program to support implementation in four BOCES (Board of Cooperative Educational Services), groups of small rural districts in CO that are pooling their resources as a way to build capacity on the evaluation system. The CEI grant provided training on the evaluation system, calibrating on the rubric, and the creation of teacher leaders to do peer coaching.

In addition to training and certification for evaluators on the front end, it is also important for SEAs and districts to monitor the evaluation results on the back end by looking to see if evaluators are achieving a meaningful distribution of observational scores and how well-aligned those scores are with student achievement data. But even in states that centralize evaluator training there is variance in whether or not trainees are "certified" to ascertain their readiness to conduct high-quality observations and ensure inter-rater reliability. According to Sandi Jacobs, Vice-President and Managing Director of State Policy at NCTQ, only 17 states currently require evaluators to be

¹⁰ Elevate Colorado can be accessed at: <https://www.mlpelevate.com/cdesignup.html>

certified. Tennessee, for example, requires all evaluators to pass a certification test (and annual re-certification tests) while many others do not. Black (MNPS) believes there are advantages in having a more centralized and standardized process for training evaluators: “I think that it’s helpful because with a state-mandated, state-driven evaluation system the messaging starts and ends with them and they’re able to disseminate better communication.” It is also clear that one shot training is not sufficient and that teachers and principals in the field need ongoing support. Most states experience a good deal of principal turnover and training new principals and those who arrive from out of state and bringing them up to speed when their previous state had a different evaluation system is particularly difficult. Gaddis from Tennessee’s Williamson County district agreed with Black (MNPS), remarking that: “The state has done a really good job with continuing to provide assistance and professional development. Our administrators have to be recertified every year and that was something that they’ve groaned about a lot, but it’s been a good thing. From a district level, it has helped us kind of maintain quality control because we know they’re all getting that state-level training.”

Sara Heyburn, Assistant Commissioner for Teachers and Leaders in the Tennessee Department of Education (TDOE), notes that they look at the data and work with the schools that have a pattern of ‘misalignment’ in terms of not showing a logical relationship between the individual growth component and the observations. Those are the schools that they target for additional support. She stated that “what we found is that it’s not really districts but more schools specifically and the evaluators within the school that may not be giving feedback that shows any relationship to the individual growth data.” Where they see at the school level a systematic pattern of misalignment, the state offers the district optional support in the form of a TEAM coach. These coaches, trained and supported in part the National Institute for Excellence in Teaching (NIET), work with the schools that accept support to go shoulder to shoulder with the principal in those schools. To diagnose what’s going on, where the data is falling short, and help the principals really begin to understand what they need to do to make sure that teachers are getting accurate and meaningful feedback.

Heyburn believes this approach has been very successful, not only in terms of closing the misalignment in the school and seeing that persist even after the TEAM coaches leave, but also because these same schools have made dramatic gains in student growth. She observes that “those schools that are receiving additional support actually outpaced the state in the last few years in terms of student growth. So while principals are not happy to be on the list initially, our 10 coaches are well-trained to ensure that their work in the schools is very much about support. She reports that over the last four years the state coaches have worked with approximately 75 schools a year and they are building out the coaching model so that schools that receive support one year, in the next year will serve as a resource for their district and region.

Tennessee’s SEA-led, centralized and uniform training and support effort appears to be the exception rather than the norm, however. In other places, there is an optional kind of state training with no test. Some states have districts really doing the training, while elsewhere (as in Colorado) the state is relying on outside evaluators that are trained by department or the regional education agencies. NCSL’s Exstrom commented that:

I think it’s a mess. I am surprised that there wouldn’t be centralized training at the state level in more places so that all evaluators meet some basic requirements. To me it seems like it would be much wiser to create a system where those who are doing the evaluations—whether it’s principals or outside observers, whoever is doing them—have to meet some sort of expectation that they know what they’re doing. And that their evaluations are going to be authentic and

meaningful and they're not going to be subjective or unreliable. I don't think that teachers are confident in places where they're not seeing that the evaluators are having to meet a minimum standard or are receiving some sort of training that absolutely ensures that they know what they're doing.

PRINCIPALS' NEW RESPONSIBILITIES

As the primary evaluators, school principals are crucial to ultimate success of these new evaluation systems.¹¹ However, finding enough time in already busy schedules for principals and teachers to do the lengthier and more numerous evaluations, to have the conversations about the results of the observations, and to find ways to use that information to modify and improve instruction is a major challenge. Some states have tried to redefine the principal's role to reallocate some responsibilities or provide some external capacity to help with that but this appears to be a work in progress. NCSL's Exstrom observed that: "The principals don't really feel like they are in a position to be doing this. They're not fully confident in their ability to do it or their understanding of exactly what this entails. They're having to spend an enormous amount of time on these evaluations and there is concern that is taking away their ability to be building leaders, to handle disciplinary issues, that sort of thing." This is particularly challenging because many schools are understaffed and had to cut vice-principals due to the recession. PA's Hardy:

If you ask the field, finding enough time is a big challenge, time to include the changes that our new system requires in addition to their everyday work. How do you carve that time out when you're still expected as a principal to make sure all the kiddos get on the bus, you know? Having it put into the school day, it's a challenge. It is a real challenge and one thing PDE can't give them is more time. What we can give them is strategy.

The Department created the Pennsylvania Inspired Leadership (PIL) Program to train principles in this work. PIL is a statewide, standards-based continuing professional education program for school and system leaders. The cohort-based program is focused on developing the capacity of leaders to improve student achievement and is run in collaboration with the Pennsylvania Intermediate Units and other partners at eight regional sites. Every principal who is new to his or her role for the first time must secure up to 150 hours of what we call PIL credit. They also have a train-the-trainer day where PDE brought their Intermediate Unit points of contact and they went through the whole principal evaluation document, all component parts of it and then the expectation is they would do turnaround training to the districts and then they would also be a conduit whereby they can provide information back to PDE. But Volkman (PDE) acknowledged that "we're changing the tire while the car is moving."

Snider (RIDE) acknowledged concern about making sure that principals have the time to do these evaluations and to do them well but she noted "that's really not something that we can solve at a state level. That's a local issue. And so we've made sure that through our bully pulpit we can frame the issues about how important this work is if we are truly committed to supporting educators' ongoing development. It takes time to be in classrooms and understand each teacher's strengths and areas that need support." She added that districts in RI are choosing to address this issue in different ways.

11 Shelby Cosner, Steven Kimball, Elizabeth Barkowski, Bradley Carl, and Curtis Jones, "Principal Roles, Work Demands and Supports Needed to Implement New Teacher Evaluation," UIC Research on Urban Education Policy Initiative, volume 3, Book 3, July 2014, Access online at http://cpre.org/sites/default/files/other/2021_ruepifinalcopyjuly2014.pdf

So in a couple of districts, people have hired a central office staffer who is going to help with all of the talent development. And that means everything from co-observing with principals or providing principals with support that they need to be able to get everyone evaluated in their building, and to provide ongoing support and professional development on the use of data. Others have created new collective bargaining agreements whereby teacher leaders can help with the observations. Snider (RIDE) said “we think that’s fabulous. So that the department chair in the science department can help with the observations of their science teachers. That way you can drill down to the really content-specific observations that teachers say they really want. While some of that work has started, I couldn’t say that at this time every district has put 1 of those 2 types of approaches in place.”

However, Snider (RIDE) noted that “We got a lot of pushback from superintendents and principals who said that our new evaluation system created an enormous capacity challenge for their building principals.” Responding to union pressure, the RI Legislature passed a law saying that highly effective teachers can only be evaluated once every 3 years and effective 10-year teachers can only be fully evaluated every 2 years. While sympathetic to the capacity challenge principals face, Snider (RIDE) feels these kinds of decisions are better left to regulation and the discretion of administrators than to be codified in legislation which limits the discretion of management. Colorado has addressed the problem of inadequate time for principals to do evaluations by establishing a process whereby a non-principal can be trained to become an approved evaluation provider. Colorado’s evaluation law does not require a principal’s license to be an evaluator but they have to be trained on the system. And once trained by CDE, a district can declare itself a peer training provider to help provide capacity for other districts.

USING THE NEW EVALUATION DATA

To date much of the effort and scrutiny around teacher evaluation has focused on getting the system operational but the next stage of implementation—and the one that will be crucial to the long-term impact of the new systems—will be focusing on how to use the new information that is gathered to better guide personnel decisions and instructional improvement.¹² Ensuring that the new evaluation systems function as designed and generate reliable, high-quality data does not in and of itself ensure that the data will be put to good use by teachers and administrators. As Weisberg (TNTP) observed,

In many districts, it hasn’t been a core priority for assistant principals, principals, people who manage principals, to constantly assess the quality of instruction and then to help teachers actually improve. It’s one thing to provide feedback to a teacher that she isn’t meeting the standards of a new evaluation system regarding student engagement, but the first question the teacher is going to justifiably ask is, ‘What can I do to get better at that?’ And there are many administrators who aren’t used to having an effective answer to that question. So, the successful implementation of a evaluation system requires—resources, such as guidance documents and trained support staff but also ensuring that evaluators have the skill and ability, and that’s really quite different.

Snider (RIDE) added that:

The point of evaluation is not just about getting it done to have summative ratings. It’s about

¹² Kathryn Doherty and Sandi Jacobs, “Connect the Dots: Using Evaluations of Teacher Effectiveness to Inform Policy and Practice,” *National Council on Teacher Quality*, October 2013.

having information that's helpful and actionable. So I would say we are at the nascent stage of that work. We held regional meetings with superintendents in groups and shared their data with them, raised some questions, talked about potential uses of the data. And while they're very interested in the promise of the work, I don't think they feel comfortable yet to actually make those kinds of practices inform their human capital decisions. . So I think they need a lot more support in that area and they need to feel confident that their data are accurate reflections of practice, and that's something we're working on—it's part of our long-term vision.” Jacobs from NCTQ acknowledged this challenge too: “we've had a system that only forced those kinds of conversations in the most egregious situations and now we're asking principals to be critical of good teachers and not just of teachers who really need improvement. To tell the person who's been in the classroom for 15 years and never received feedback or a negative evaluation that she has areas of weakness—it's as hard to hear as it is to deliver. There are a lot of human dynamics at play here.

Tennessee's Heyburn concurred with this assessment noting that:

I think people have really seen the value in evaluation and support systems over the last few years in terms of driving the feedback process and that has translated into improved student results statewide. I think it's less of a hearts and minds issue now and more of a skills and capacity issue. Evaluators might be accurately scoring, but they don't always know how to take the next step and turn that into meaningful feedback for teachers. Or how to set up schools so that teachers are more engaged with each other to leverage strengths and address challenges that you can see through the teacher evaluation system.

Ironically, however, it appears—at least in the short term—that the new teacher evaluation systems may actually deliver less instructional coaching than existed under the old systems because of the increased time required for principals to do the extra observations. Poda (CCSSO) observed that:

A lot of principals are frustrated with some of the requirements of implementing evaluation at the expense of ongoing feedback to teachers. The day to day instructional supervision has kind of gone by the wayside because of the effort to make sure principals are observing every teacher three times for evaluation purposes. That brings up the issue of who else can help with evaluations and how can schools and districts get the resources to be able to hire those people or expand the roles of teacher leaders and assistant principals, district office staff and others who can assist with instructional improvement as well as evaluations. Otherwise, principals will be really overwhelmed and unable to provide the instructional leadership today's achievement expectations are demanding.

In this regard, it is important that the principal evaluation system be aligned with the new teacher evaluation system to ensure that principals appropriately prioritize and are incentivized to assess and coach teachers with rigor and objectivity.¹³ In Pennsylvania, for example, their new principal evaluation piece “The Framework for Leadership” followed a year after the introduction of the new teacher evaluation system. Volkman notes that “the Department developed what we call a ‘connectedness document’ to show that there are correlations between a framework for teaching and a framework for leadership. We actually brought practitioners to the table and sat down and we

13 Patrick Schuerman, et al., “Supporting Principals to Use Teacher Effectiveness Data for Talent Management Decisions,” May 2014, Accessed on June 25, 2014 at <http://www.principaldatause.org/assets/files/reports/Findings-Recommendations-201405.pdf>

developed a set of guiding questions that principals can use with teachers.” They have done this both horizontally and vertically because they’ve also developed a set of guiding questions for principals to have with superintendents and supervisors so that they’re all on the same page. Hardy:

We had this long runway and it helped folks out in the field understand, internalize, and then apply what came out of the implementation phase. Our vision was to involve practitioners every step of the way. Although the design and the responsibilities were at the State level, the development and the tools and the understanding of what was needed came from the field and I think that is so critical.

Tennessee also redesigned its principal evaluation system to better align with its new teacher evaluation system. Heyburn stated that:

While we were also implementing principal evaluations from year one, per our state law, the model that we were using needed to be updated. It didn’t give enough priority to what was emerging as the core competencies that principals need to be effective in this new landscape that includes intensive teacher evaluation, and transitioning to Common Core State Standards. We had to make sure that we had systems that were aligned in terms of focusing on teacher development and teacher support...getting the school leadership piece right is so huge when you think about teacher development, because those are the people who are gatekeepers for teacher feedback and support.

The state piloted significant changes to its principal evaluation rubric in 2013-2014 for the statewide roll out in 2014-15.

Many folks working out in the field drew a distinction between structural challenges and cultural challenges in the implementation work. The structural issues involve finding time for principals to do the evaluations, setting up the technology platforms to collect data that is accurate and secure and informative, and providing professional development to implement the assessment system with fidelity use the data to make decisions. Snider (RIDE) remarked that:

There’s also a whole host of cultural challenges that continue to need more thinking and discussion. One of them is even though you can get a group of evaluators to watch video and calibrate their thinking in fairly reliable ways in a training setting, it becomes quite different when you are in classroom observing a teacher that you might have known for 5 or 10 years and you have lunch with. It requires a cultural shift for a principal to feel comfortable giving feedback, both positive and areas for growth, I think everyone’s feeling some discomfort taking on that role.

Weisberg (TNTP) concurred, noting that:

You’re talking about shifting not just systems but very deeply ingrained cultural attitudes against treating teachers as individual professionals with their own strengths and weaknesses and that shift is not going to happen overnight...some people are inclined to see this as the last chapter when it’s really the first chapter. But we advocates didn’t do a great job of setting realistic expectations and talking about the magnitude of the cultural shifts and that it would take multiple years to really see momentum being built.

Another cultural shift—and not just within the education community—is to change the perceptions around the purpose of the new evaluation systems. First, it is necessary to move the conversation from the punitive—focusing on getting rid of a small number of bad teachers—to the productive—using better information to improve the instruction of all teachers and creating a continuous improvement model.¹⁴ In most states—even those that have already implemented new evaluation systems—the vast majority (and often upwards of 95%) of teachers are rated as being highly effective or effective, so it can be hard to communicate that it’s worth doing evaluation. Snider (RIDE) believes that “the message needs to be more clearly articulated that the new evaluation system is a process whereby we can provide even our effective and highly effective teachers with really good feedback and tailor their professional development and learning goals to be exactly what they need in more precise ways.” States will also need to work with the universities that run their teacher and principal preparation programs to ensure that the purpose and content of the new evaluation systems are communicated even before these folks enter the profession.

For evaluations to be used to inform classroom instruction, teachers need to be provided with differentiated, targeted professional development that can accommodate the wide range of academic disciplines, grade levels, student demographics, and instructional specialists (i.e. ESL and special education teachers). But as Hardy (PDE) observed, “the challenge is trying to provide professional development to all the different key groups with the fidelity that the state has intended so that we don’t get the ‘whisper down the lane’ and we don’t send out mixed messages. Making that professional development available to everyone and without cost to those people to attend is very difficult.” PA has a statewide system of support in which the SEA provides the funding for the Intermediate Units to provide the training. They also have statewide professional development and a staff portal, which is free and designed to provide resources that any educator can go to when they need professional development or want to join a professional learning community.

DATA SYSTEMS, REPORTING LOGISTICS, AND MONITORING

A crucial piece of infrastructure for the new evaluation systems is the data collection and reporting systems that districts and states use to gather, analyze, and disseminate the information collection about teacher performance: observations, student surveys, and student growth scores. This is a place where scale is helpful and the state can come up with solutions that are going to be more efficient and reliable than each district dealing with it on its own. Weisberg (TNTP) argues that:

It’s very important to have people whose full-time job it is to maintain the data system, which can be a huge pitfall. Just having a place where educators who are out in the field can enter the evaluation data and be able to access the data and analyze the data and being able to make sure people are getting questions answered and problems are getting trouble-shot. I mean this is one of these issues where it’s really go big or go home. In other words, if you are going to try to do this on the cheap, you’re going to end up with a bad result.

Snider (RIDE) believes their state data system “has been a really big success for us.” Using RTTT funds, they developed the Educator Performance and Support System, a platform that helps districts

¹⁴ Stephanie Fagnani, “Drilling Down on Education Data,” *District Administration*, Accessed on October 1, 2014, <http://www.districtadministration.com/article/drilling-down-education-data>

schedule evaluations, collect data, and provides evaluators with all of their data at an individual and aggregate level. Snider (RIDE) notes that:

It's also a way for districts to upload summative ratings to send to the state at the end of the year, because we have connected evaluation data with our certification system. And that's working really, really well. Without that it would make all the capacity issues that much larger and it would make it more difficult for us to use data in all kinds of ways we do currently and plan to do, whether it's for improvement or certification or any number of human capital decisions.

Colorado has also built a management performance platform to help districts manage the data from their evaluations. Use of the system is free but optional for districts in the state and the system was scheduled to be operational for the 2014-2015 school year. The platform contains four modules: the first it is a guide to scoring the rubric for the evaluator. The second module is the 50% of the student growth that's measured on the evaluation. The third module aggregates the professional growth measures for students and the professional practices for teachers to get a final rating (ineffective, partially effective, effective, or highly effective.) The fourth module has the data that tells teachers, districts and principals where teachers are doing well, where they're not doing so well, and how they can create some individualized PD and professional learning to spur improvement in practice.

How states organize and disseminate final evaluation information back to teachers and administrator is also important but varies widely. One key decision for states and districts—and another area where there is considerable variation—is to determine how to align scores on new evaluations with each level of performance. Thanks in part to a collaborative partnership with the Pittsburgh Federation of Teachers and significant investment from private funders including the Bill & Melinda Gates Foundation, Pittsburgh Public Schools (PPS) was able to develop their evaluation system for several years before evaluative stakes were attached. The District used this time to inform decisions like setting performance ranges, and educate and engage teachers around its use. In Pittsburgh's system, teachers ultimately earn between zero and 300 points during the whole evaluation process that includes observation, student growth, and student survey tools.

Sam Franklin, the Executive Director of the Office of Teacher Effectiveness for the PPS, emphasized the desirability of using several years of evaluation to inform this decision.

We had the benefit of being able to decide on performance ranges based on multiple years of real evaluation results for real teachers in Pittsburgh Public Schools. So with a large and fairly mature set of data, we were able to look at that data and hone in on where to set the standard for teacher performance in a way that took into account the policy context and aligned to our goals for students.

In Pittsburgh, teachers who earn less than 140 of 300 points perform at the lowest level (Failing) in the evaluation system. To perform at the highest level (Distinguished) teachers must earn at least 210 points. In most cases, to be dismissed based on performance teachers have to perform at the lowest level in two consecutive years. Franklin (PPS) says that these performance levels represent real differences that matter for students. "Multiple measures are supporting the fact that these classrooms are different from each other in a way that makes a difference for students. At the same time, it's important to not create a quota system. The system is designed in a way such that all teachers could perform at the proficient or distinguished levels, if their performance merited that result."

There are also lots of concerns—among both parents and teachers—about data privacy and what all of this data from student tests and teacher evaluations will be used for. The NCSL’s Exstrom remarked:

There’s a huge concern around how that information is going to flow back and whether or not it’s going to be released to the public, whether or not it’s going to be used for who knows what. And a major challenge right now is some of the legislation that was introduced this year in state legislatures to protect student data privacy, potentially could have a significant impact on states’ ability to continue with their longitudinal data systems the way that they are designed right now.

She notes that a majority of states have now seen data privacy legislation introduced that would undo the ability for states to collect information in certain ways, or even if the state can collect the information, prevent them from sharing it across different agencies. Exstrom (NCSL) pointed in particular to restrictive legislation that passed in Louisiana and New Hampshire that she believes will prevent them from continuing to use the evaluations and the information that they’ve been getting about teacher effectiveness to inform teacher preparation programs. “The big effect,” she notes, “is going to be whether the agencies can share information, so the Department of Higher Ed cannot share information with K-12, K-12 cannot share information with Early Ed, so those data linkages will be undone.” Even if new evaluation systems are implemented successfully and generate reliable and useful information, such limitations on data usage and sharing could potentially preclude policymakers and administrators from using the information in the ways it was originally intended.¹⁵

In addition, while there is a lot of talk still about the changing SEA role in moving from being a compliance monitoring organization to more of a school improvement and support services unit, states are still figuring out how to do both simultaneously. The monitoring role is still necessary—and required—both as condition of federal grants but also to ensure that districts are doing this work the right way and in accordance with federal and state laws and regulations. Exstrom (NCSL) remarked that:

This is just something that I think state policymakers hadn’t thought about until we hit this place where we’re supposed to now be moving forward. The SEAs don’t necessarily feel comfortable or exactly know how to monitor it. They don’t know if they are empowered to monitor it. And they don’t know how to report to the Feds that they are monitoring it. They don’t know if they can confidently report that they were monitoring and they are empowered to monitor.

NON-TESTED TEACHERS AND STUDENT LEARNING OBJECTIVES

A majority of teachers work in untested grades or subjects and figuring out how to measure student achievement or growth in their classrooms remains perhaps the biggest problem in the implementation of new teacher evaluation systems, which was highlighted in a recent Reform Support Network report.¹⁶ Districts have addressed this challenge in a variety of different ways and with widely divergent levels of effort.¹⁷ Some districts are relying on SLOs—student learning objectives—or

15 Susan Dynarski, “When Guarding Student Data Endangers Valuable Research,” *New York Times*, June 13, 2015.

16 Government Accountability Office, “Race to the Top: States Implementing Teacher and Principal Evaluation Systems Despite Challenges,” September 2013.

17 Reform Support Network, “The View From the States: A Brief on Non-Tested Grades and Subjects,” accessed July 14, 2014, <http://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/view-from-states.pdf>

SGOs—student growth objectives—that are designed by teachers in partnership with administrators. But not nearly enough is known about SLOs and SGOs and there are concerns about how objective, reliable, and comparable these are as measures of teacher performance. Other districts are relying on portfolios of student work, especially in the arts, but like SLOs portfolios are criticized for their lack of standardization and comparability. And some states are recommending that districts use scores from commercially available achievement tests as a measure of teacher effectiveness. Jacobs (NCTQ) noted that:

Places have gone in so many different directions here—places looking for an actual assessment to use for every grade and subject, versus places just using SLOs, versus places using less standardized measures and portfolios. In many places, I have seen that teachers may lack the data literacy and the assessment literacy to really be able to write an SLO. Data literacy and assessment literacy are not something that teachers get a lot of training in and I think that we're seeing that now as it plays out around SLOs. I think in some places it's not that teachers don't know how to assess their students; it's that they don't have the vocabulary to be able to talk about it and write about it and know how they would then set an objective for themselves that matches it. But I think in other places we're seeing there may not really be much of a plan in place to measure whether students in that particular course are learning anything. Then there's the piece of it where teachers are pretty afraid of this new system. They don't really understand it. It hasn't been around to build any trust yet and now they're being asked to set up targets for themselves that they are seeing very high stakes attached to. That does not sound like a recipe that's going to motivate teachers to set a rigorous target.

Most states appear to be relying on principals to be the gatekeeper here but it isn't clear that they really know what a good SLO would look like or what an appropriate target is either. On a positive note, over the next couple of years we should start to see some data that show us how different approaches compare to each other and that will give the field some good information about how to move forward on this issue. Snider (RIDE) admitted that:

To be perfectly candid, this continues to be a work in progress because the process of having really rigorous and well-calibrated SLOs demands that a district has a strong comprehensive assessment system and educators who have assessment literacy. And that doesn't mean that they just purchased a good number of tests. We're really encouraging districts to be less dependent upon commercially available assessments, but to do more work within grade levels or departments to create curriculum-embedded assessments that are reflective of what has been taught and that really push students to do problem solving and higher level, more rigorous work, as a way to measure their own growth and learning of the curriculum and also to inform SLOs. But that's a hard body of work, and it certainly can't be accomplished at every grade level in every content area in a couple of years.

Some districts—like New York City, and with funding from the Gates Foundation, Hillsborough, Florida—are designing their own assessments from scratch but few districts—and particularly few small districts—have the resources or capacity to undertake this kind of work on their own. As a result, Weisberg (TNTP) believes that SEAs can play a very productive role in identifying and/or designing assessments for districts to use. He stated that “I think SEAs should be at a minimum, a clearinghouse for quality assessments that are aligned with state learning standards. I think beyond that they should probably be generating optional assessments districts in subjects outside of English and math because a lot of districts aren't going to have the resources or expertise to design their own

high-quality assessments.”

Tennessee appears to be an excellent model of how an SEA can assist districts with this work, as the Department has developed a variety of alternative growth measures that are optional for districts to use, including in the areas of world languages, physical education, health, fine arts, special education, and pre-K and kindergarten.¹⁸ Heyburn estimates that about 50% of the state’s teacher population is generating individual growth scores currently and that if districts used all the measures that are currently available that number would be closer to 70-75%. Nashville’s Black (MNPS) applauded the work of the TDOE in this area, noting that the state has created a fine arts portfolio model and that for all fine arts teachers there is a portfolio that’s created throughout the year, kind of pre- and post- testing. That portfolio, towards the end of the year in about March or April, is given a blind one-to-five score from other fine arts teachers across the state which then becomes their individual growth measure. Heyburn noted, however, that:

There is a group of teachers who may never have individual growth measures based on an assessment that they would give to their students. For example, media specialists have indicated using the school literacy scores is an appropriate measure of student growth for their evaluation. We don’t want to require a state student assessment that is purely for the purpose of teacher evaluation. Any test that we require, or even make optional for districts, should be telling us something valuable about student learning. We’re trying to balance the tension between not over-testing students, but also making sure that teachers are getting individual, actionable data through the evaluation process.

Gaddis (Williamson County, TN) cautioned, however, that a large number of teachers are still not being evaluated based on student growth in Tennessee. He estimated that 55% of their teachers don’t have value-add scores, including their K, 1 and 2 teachers and all of the elective teachers. As a result, like teachers in many other districts across the country, they are still stuck using the school-wide composite score for a large number of folks and many feel that’s just not really indicative of what’s happening in their classroom. Black (MNPS) observed that:

The non-tested piece is a national question. It is a Tennessee question. It is a Metro Nashville question and teachers and principals have said that it presents a problem in two ways. One, sometimes great teachers don’t get an overall evaluation score that reflects that they’re effective teachers because their school value-add can be low. We also hear there are teachers in the building that aren’t what most would deem effective but their overall score reflects that they’re more effective than an observation score or professionalism rubric might show because they are able to ‘hide’ with a good school-wide growth score.

Black (MNPS) praised this approach, remarking that “We’ve been very fortunate over the past couple of years to have NIET (National Institute for Excellent Teachers) coaches support many of our schools in this way. They provide constant on the ground support for principals. They assist with rubric calibration and norming as well as providing principals with feedback on their pre and post conference delivery. Last year I believe we had ten schools that were offered this kind of support and it helps achieve a more normal distribution of scores especially on the observation side of the evaluation.” Poda (CCSSO) added that:

¹⁸ For more on Tennessee’s efforts see: “Teacher Evaluation in Tennessee: A Report on Year Two Implementation,” Tennessee Department of Education, accessed from http://www.tn.gov/education/teaching/docs/yr_2_tchr_eval_rpt.pdf

People don't feel great about the SLOs. They see SLOs as a way to get student growth data for teachers whose students don't take state level assessments but I think they still have questions about comparability and reliability. In other words, how one person reviews the goals or the objectives that a teacher puts forward versus someone else in a different school or another district; and whether it's an equitable across schools and districts. And in places where they don't have SLOs, districts have had to rely on using things like a whole school's data rather than something specific to their subject area. Subject area specialists are scratching their heads and saying, 'Now how am I responsible for Miss Jones' classroom of students?' We're still trying to figure out the best way to measure student growth in all subject areas.

There is wide variation among states in how far along they are in creating SLOs and aligned measures and in terms of how centralized the process is. Pennsylvania piloted a voluntary SLO process in 2013-14 that was mandated in 2014-2015. Hardy noted that:

It's been a challenge for us because it's new learning and it is an area that many educators as well as trainers have not had the opportunity to spend as much time as they have with observations and teacher practice. So it's a little bit of a heavy lift. But I'm going to be honest. I don't know if we can ensure the quality of SLOs because schools are not turning their SLOs in to the State, to PDE to have them approved.

The PDE worked with an expert to design training, resources, templates, and to train their Intermediate Units, train their trainers, and pilot. They are in the process of vetting the models that came from the pilot and will provide the exemplars and supporting resources to districts free of charge. Hardy remarked: "That's the best we can do because we cannot be in 500 districts and looking at this. So, it is a challenge and we don't have any data yet because we haven't implemented it."

In New Jersey, Shulman (NJDOE) observed that:

Getting practitioners talking about measuring: measuring academic performance, measuring the growth of their students, using data to inform the discussion to determine what sort of outcomes they want to see, that brings the inherent benefit. Now with that being said, we know that there are districts, schools, and individuals that are struggling in the first year to identify what the appropriate measurement tool or assessment might be to measure student growth, what an appropriate threshold for growth might be. How do you collect all the data? How do you use it to refine what an appropriate sort of bench mark is? Now even the folks at our NJEA, which is our largest teachers union here, has said that in general, the SGO conversations are going well despite implementation challenges. But in pockets where they are not, we've put in an appeals process.

Tim Matheny (NJDOE) added:

Creating quantifiable goals for how much a student learns over a course of a school year is really a new task for many, many classroom teachers. And we have provided lots and lots of support around that and moving forward, we have a lot of confidence that student growth objectives are meaningful and teachers are increasing their capacity to set good targets on really important aspects of their instruction. One of our mantras is that SGOs should be student-focused, teacher-driven and administrator-supported.

In the spring of 2014, the NJDOE conducted 39 SGO workshops and it reports that more than 25,000

educators attended some form of training, workshop, or presentation offered by the Department in 2013-14. The Department's analysis of SGOs in use in the field found that 70% had high-quality, specific, and measurable statements for student learning but that 30% lacked specificity and that educators were inconsistent in how clearly they connected SGOs to specific standards.¹⁹ Most districts across the country appear to remain far from achieving the goal of incorporating student academic growth into teacher evaluations. An August 2014 Bellwether study concluded that despite state policy changes, "many states and districts aren't embedding student growth in evaluation ratings in any meaningful way."²⁰ Clearly there is much more work to be done in this crucial area.

BALANCING MANDATES/ STANDARDIZATION AND FLEXIBILITY

A continuing challenge for states (as well as for the U.S. Department of Education) is to determine what parts of the evaluation process should be regulated and standardized and in what parts districts can be accorded greater flexibility. This is most apparent in the different approaches that states are taking to the kinds of observation instruments that they permit districts to use. New Jersey's Shulman noted the importance of:

Finding the balance between local control and state policy. We made a conscious decision to have a broad variety of evaluation instruments for districts to choose from. We have somewhere around 20 or so instruments that folks can choose from as opposed to one size fits all instruments. We thought about the context of New Jersey and said for our state, the diversity that we have, the history of local control, the differences that we see in these districts, we wanted to provide those choice options.

Tim Gaddis (Williamson County, TN) observed that a big change in Tennessee has been that the state now gives districts an annual opportunity to apply for flexibility around a variety of different things such as the number of observations, the frequency of the observations, and the scoring. One such request that they asked for, and which was granted, was to change from a snapshot kind of approach to more of a cumulative approach to grading the teachers. He noted that there are certain items on the rubric that aren't going to be present in every lesson, such as an indicator on grouping for instance.

So what we did is we—we did the same number of observations – well a few more actually – but we don't record within the system the final rating after each observation. We wait and do a single rating at the end of the semester on each indicator. So that allows us several times visiting to see if there really is evidence of that particular indicator in the classroom. That was huge for us. It built trust in our teachers and they didn't feel as though it's a 'gotcha.' So they're able to really work with the administrators and it's more of a coaching and working together kind of a relationship.

At the end of each school year the TDOE looks at the observational scores for all of the teachers in a district and compare that to their students' growth scores to ensure there is not a large mismatch. If

19 New Jersey Department of Education, "2013-2014 Preliminary Implementation Report on Teacher Evaluation," available at: <http://nj.gov/education/AchieveNJ/resources/13-14preliminaryteacherevalreport.pdf>

20 Chad Aldeman and Carolyn Chuong, "Teacher Evaluations in an Era of Rapid Change: From "Unsatisfactory" to "Needs Improvement," Bellwether Education Partners, August 2014, accessed September 15, 2014 from <http://bellwethereducation.org/publication/teacher-evaluations-era-rapid-change>

the numbers are more than two off a five-point scale on a significant number of the teachers, then that district loses their flexibility and the TDOE provides on-the-ground assistance from NIET trained TEAM coaches to those particular schools. Black (MNPS) praised this approach, remarking that:

We've been very fortunate over the past couple of years to have TEAM coaches support many of our schools in this way. They provide constant on the ground support for principals. They assist with rubric calibration and norming as well as providing principals with feedback on their pre- and post-conference delivery. Last year I believe we had 10 schools that were offered this kind of support and it helps achieve a more normal distribution of scores especially on the observation side of the evaluation.

Interestingly, Tennessee also enacted the High Performing Districts Act which allows a small number of districts with above-average student achievement to request additional regulatory flexibility from the Commissioner in teacher evaluation and other areas.

States are taking different approaches to monitoring and regulation, based in part on the relative strength of local control and the historical role of the SEA in their particular state. Some states have legislated and regulated much of the process of teacher evaluation leaving little discretion in the hands of evaluators while other states have adopted more of a “trust-but-verify” kind of approach that relies more on back-end monitoring. Katy Anthes, the Executive Director of Educator Effectiveness at the CDE, for example, remarked that:

So we have around 65 to 100 metrics that we're going to be analyzing once we get all of the data. And what we're looking for there are sort of outliers or anomalies in the data where we think we want to have a conversation with the district. So, if for example, a district is rating all of their educators exemplary but yet their student growth trajectory is going down, that would be a district we would want to have a conversation with around the rigor of how they're doing the evaluation systems. We know it's going to take them a few years to do this work really well and we really see our role as a support role to help them do the work in a meaningful way.

Another issue centers on how SEAs should support and leverage the work that particularly innovative districts—often those like Pittsburgh that started the teacher evaluation reform work in advance of the PDE (in this case with help from a Gates Millennium grant.) Language in the PA statute says that districts can apply to have an alternative evaluation system from the state model as long as it is as rigorous and meets requirements from PDE. Pittsburgh's application was recently renewed, but not without controversy over the opposition of the city's teachers union.

A related issue concerns the desire to avoid teacher evaluation reform from becoming a compliance exercise where districts or school leaders merely go through the motions rather than truly engaging and investing in the work. Weisberg (TNTP) noted that:

There are certainly districts that looked at this more as a compliance exercise than an opportunity to really create much richer discussions about instruction and a better focus on the quality of instruction and on real development of teachers. They just saw it as a compliance exercise and so that's what they got; an elaborate box-checking exercise. There are districts that did not allocate sufficient resources to the implementation effort. There are districts that spent much more time than they should have on the design of the evaluation system and not enough time and resources on how the system would actually be implemented.

SEAs and districts that have adopted a partnership mentality around the evaluation work that involves

frequent communication seem to be enjoying the greatest success. Black (MNPS) noted that “part of the success of implementation for MNPS is the relationship that we have with the TDOE and their willingness to collaborate around teacher evaluation. We communicate very regularly on questions that I have and they are very responsive.”

INTER-STATE AND INTER-AGENCY COMMUNICATION AND THE “SILO” PROBLEM

While there is tremendous variation in the design of states’ evaluation systems and how far along they are in implementation, it is imperative that they share information and learn from one another to identify effective practice. States that won RTTT grants report that the Reform Support Network run out of the U.S. Department of Education’s Implementation and Support Unit was a valuable forum for states to share lessons from the field. Snider (RIDE) remarked that:

There is a lot of information sharing among us. There has been considerable learning because of what has worked well and where there have been struggles. We’ve talked a lot about how to communicate with the field to make this an iterative process. We’ve talked about ‘How do you factor in student learning in an evaluation system in a responsible way’ and try to learn from each other about who is doing that well. We don’t take advantage of cross-district sharing and we are trying to set up the conditions to have that occur more regularly.

Similarly, given the interconnectedness of teacher evaluations with standards, assessment, and curriculum it is important that the administrative units that manage these different areas in SEAs and districts align their respective efforts. In RI, for example, Snider (RIDE) reports that:

The Educator Evaluator unit is making really deliberate connections among offices to make sure that that work is well-coordinated and understood by staff who might not see themselves, for instance, as having a direct hand in educator evaluation. But all the work they’re doing with the district on curriculum development aligns with the Common Core in helping them build local assessments; it’s always connected to SLO work, and we make those connections really visible to people as part of the work.

This kind of communication and coordination is crucial given the long-standing concerns about “siloeing” in SEAs.

NCSL’s Exstrom cautioned that:

There’s been a huge challenge with buy-in. In many states, unfortunately, the policymakers themselves still are not communicating in the way that they probably should, and that’s the state boards, the SEAs, the legislators, the governors, in many cases they’re just doing their own separate piece of work but they’re not really informing each other about how things are going. I think in many cases, there’s been a big challenge with communicating this to teachers. I’m still surprised as I go around and talk with teachers from across the country at how confused and how much misunderstanding there is about the evaluation systems and where this came from and how it’s being designed and whether or not they truly have a voice in it and that sort of

thing. The lack of communication has really hampered implementation.

At the district level, funding cuts due to the economic downturn have hurt and the challenge is trying to leverage the staff that they do have in the most efficient way to meet their needs moving forward. Kirby (Harrisburg Training and Technical Assistance Network in PA), notes that as a result they:

Need to think differently about those school-based teams and district-level administration and their roles and functions and their cross-training roles and functions. In some of our larger districts, one of the struggles they have is at the district office they have staff that are assigned specific jobs--someone's responsible for Title I, someone's responsible for Title III, someone's responsible for special education, someone's responsible for personnel. Because those roles have been solely defined as such, some people have struggled with how to allocate and/or cross-train and have folks doing different work. And districts are asking for help in how to think differently and more strategically and how to do that.

Effective vertical communication—in both directions—between SEAs and districts and schools is crucial and that often depends on building relationships and developing trust. Black (MNPS) noted that in Nashville:

One of the things that we've really tried to do is to make teacher evaluation (policies and procedures) in MNPS transparent for all stakeholders. So part of that is writing standard operating procedures which had not been done prior to my position. Those are published and publicly available through our Human Capital website so it's forward-facing and referenced to teachers, principals, and central office staff on an ongoing basis. We have also held focus groups with teachers. We did that with about 100 teachers this year to learn what's going well? What's not going so well? We talked about things that are in local control, things that are state-controlled and brainstorming what could be some potential solutions. We drafted recommendations at the district level and then recommendations for the state as well to improve the processes around teacher evaluation.

POLITICS, LAWSUITS, AND FEDERAL WAIVERS ARE COMPLICATING IMPLEMENTATION

While there is no question that the U.S Department of Education was able to push states forward on teacher evaluation reform through RTTT and the NCLB waiver process, some observers have questioned the commitment of states that seem to have pledged to create new evaluation systems primarily as a means to get a grant or a waiver. Sandi Jacobs (NCTQ) notes that:

I personally worry about the states that are trying to move forward primarily based on a waiver promise rather than some underlying state policy—finding both monetary resources and leadership resources and everything you need to do this well, it's hard. There are certainly states that applied in their waiver just pointing out what they already had in State policy and, so, I think those states are in a pretty different boat than the states that really have no underlying policy at all—whether through the legislature or the State board or anybody else saying, 'Here's how it's going to look in this state.' Most of the waiver language is very broad and very vague,

so how you get from there to real implementation I think is quite a challenge. There are states where what they appear to be promising in their waiver—especially using student achievement in a significant way—then doesn't seem to line up that closely with what they're actually doing. That's where Washington found itself and I think other states like Oregon and Kansas are in a similar boat there. It's a headache for the Feds and it's certainly a problem for those states.

Opposition to new teacher evaluation systems has mounted in the past two years both because of concerns about their validity and fairness to teachers and because the new evaluations are tied to new Common Core-aligned assessments. This has led to considerable pressure to delay the use of test scores in the evaluations or to postpone using the new evaluations in making personnel decisions. In Louisiana, Governor Bobby Jindal has reversed his earlier support for Common Core, tried to halt the implementation of new assessments, and filed a lawsuit against the federal government claiming that it is federalizing education policy. The RIDE has postponed—from 2014 to 2017—the use of student growth scores as a formal piece of evidence in an individual teacher's summative rating. Snider (RIDE) noted that “that was in response to the field's concern about the Common Core transition. So that was something very concrete that we could do to help allay those fears.” Lawsuits have challenged new teacher evaluation systems in several states, including New Mexico, Colorado, Tennessee, and Florida. Colorado's legislature made the decision to allow districts to postpone the introduction of measures of student learning in evaluations for a year. While the technical challenges of implementing new evaluation systems are enormous, the political challenges are also large. The standard for performance established in Pittsburgh, for example, is higher than that across the rest of the state, where teachers must earn below 0.49 out of 3.00 points to perform at the Failing Level. Pittsburgh's decision to stand behind this higher standard created controversy which was covered in the local media and strained the district's collaborative relationship with the teachers' union. It is imperative for SEA and district staff to communicate regularly and clearly with educators, administrators, and the general public about the structure and purpose of the new evaluation systems. Franklin (PPS) stated that “there are challenges related to everyone in the district knowing the facts and the context and sorting out the truth from what's not true. When there is conflict or controversy or disagreement it does exacerbate what already is a big challenge—which is just communicating accurately pretty complex and detailed information about the way that the evaluation system works.”

Building educator support and buy-in with the new evaluation system is crucial, and teachers need to feel like the intention of the new system is to improve teaching and learning. If the system is seen as unfair or unhelpful—or as primarily about firing teachers—then a serious credibility issue can emerge. Anthes (CDE) remarked that:

Our engagement with districts has gone well. I feel like most districts in Colorado really feel like their department is a support to them on educator effectiveness work. So they call us, they engage with us, they say, we need training, it's a two-way communication. That hasn't always been the case with state departments and districts. I think it was over our three-year process of getting pilot districts on board going around the state training all the districts. It was a very customer service-oriented approach and messaging was around we're not all going to be perfect at this. We're going to have to learn together. We're going to make mistakes, we're going to learn from that together and we're going to get better over time. And I think that puts districts at ease and they feel like the department is not only the compliance officer.

Nationwide, however, all of this political and legal fighting around the Common Core has created great uncertainty about the future of the standards and aligned tests upon which most states' new

teacher evaluation systems are based. This environment of uncertainty adds an additional level of difficulty to already challenging implementation work. A growing backlash against testing also led to an “opt-out” movement by parents across the country during 2014-2015 and the ultimate size and impact of this movement on testing systems is unknown. The political uncertainty at the state level has been exacerbated by the wavering of the U.S. Department of Education on teacher evaluation in RTTT and the NCLB waiver process. Initially, the U.S. Department of Education denied states’ requests to delay the implementation of new teacher evaluation systems or to postpone the use of student test scores in teacher ratings but reversed course in August 2014. Some states (such as Michigan and Hawaii) have chosen to take advantage of the flexibility and extend their implementation timelines while others have not.²¹ Many states welcome the opportunity to delay teacher accountability for student performance as a way both to reduce the growing political opposition to evaluation reform and to focus on getting the implementation of new Common Core-aligned assessments right before they are connected to personnel decisions. The fact that a teacher-evaluation mandate is not included in either of the ESEA re-authorization bills that passed the House and Senate further muddles the future of these new evaluation systems.

ALIGNING THE TEACHER EVALUATION WORK WITH THE ROLLOUT OF COMMON CORE AND ASSESSMENTS

Implementing new teacher evaluation systems would be a major undertaking on its own but most SEAs and districts are simultaneously rolling out the new Common Core academic standards and the new aligned assessments. This further strains SEA and district capacity and emphasizes the need to think carefully about rolling out the new evaluation system in a logical way so that it is sequenced with the introduction of other interconnected reforms. NJ’s Shulman (NJ DOE) observed that:

We are one of probably a handful of states that are doing the following four things: a) have adopted Common Core; b) are implementing a new Common Core aligned assessment, in this case PARCC (Partnership for Assessment of Readiness for College and Careers); c) have an educator evaluation system that has some level of weighting of student growth based on those assessments; d) and then based upon the 2012 teacher TEACHNJ Act, have specific detailed ramifications from what evaluations of inefficiency lead to, or persistent evaluation of inefficiency lead to. Those four pieces I would say are foundational for where we think New Jersey is headed.

Other states, however, did not have these four key components in place when they began implementing their new evaluation systems—or even if they did—dropped, modified, or delayed one of those pieces significantly. In Tennessee (one of the earliest and most effective implementers of new teacher evaluation systems), for example, the legislature halted the planned roll out of the new PARCC assessment so in the 2014-15 school year the state’s educators taught to the Common Core State Standards but students and teachers were evaluated on the old state test (TCAP) which is a measure of Tennessee state standards, not Common Core State Standards. The state’s standards and assessments are therefore mis-aligned and educators feel that teachers cannot be fairly evaluated

21 Alyson Klein, “States Mixed on Waiver Leeway for Teacher Evaluations,” *Education Week*, October 1, 2014.

on the new standards with the old tests. This disconnect led to great confusion and frustration among teachers during a critical year for the implementation of the new evaluation system.²²

The ‘silo effect’ within SEAs and districts that has tended to keep the work of different units separate and disconnected represents a real challenge to ensuring coherence across different initiatives. In CO, however, the Colorado Education Initiative (CEI) is implementing a Gates Funded program in 13 “Integration Districts” across the state. These districts assign teacher leaders as “Integration Liaisons” who work to implement the new evaluations using a systems thinking approach that seeks to integrate the new academic standards, the assessment piece, and the evaluations together so they’re not separated but rather integrated together to promote student growth and student success. With the help of CDE they held four or five major professional development gatherings during the year. CEI’s Gradoz observes that they “turned out to be quite successful for us in terms of what teachers do in the classroom and what peer coaching can do in terms of their professional growth and learning. And it’s helped build capacity for the principals, because finding the time has been the biggest challenge for this whole.” Based on the results they’re building tools and resources to put on their website so they can be shared with the other districts in the state. Anthes (CDE) added that: “we created our rubrics to make sure that they align with the new instructional shift of the Colorado Academic Standards — which includes the Common Core. So there’s really alignment built in.” In PA Volkman (PDE) commented that “One of the things we’ve been very careful of is to show folks that it’s all interrelated, that this is all connected. A series of webinars were actually conducted this past year to show how educator effectiveness ties into our new standards because, again, all things focus on student achievement. So, we’ve tried to draw the corollaries there to let people know that.”

Several folks working out in the field emphasized the importance of piloting the new teacher evaluation systems in advance of “going live” with the new systems statewide. This can enable implementers to identify and resolve any problems that emerge and give teachers and principals time to adjust to the new system and their role within it. NJ’s Shulman, for example, noted that:

We’ve invested in having educators drive the implementation of these systems, and using educators to have a two-way dialogue and a commitment to a model of continuous improvement. So what we saw and learned from the first pilot informed the second pilot. And what we saw and learned in the second pilot informed our first year of implementation. And what we’re learning from this year informs our second year, and this helps drive not only the regulations that go with it, but our support, our policies around these pieces, our guidance documents, our outreach, our communications. So in New Jersey, we’ve had the advantage of sequencing these in a little bit more of a methodical manner and really having districts that have now had three, four, five years under their belt in preparing and getting ready for a new assessment and a new evaluation system.

22 Lauren Camera, “Tennessee Teachers Chafe at Common Core Uncertainty,” *Education Week*, November 25, 2014.

Conclusion

States are working hard to realign education policies, institutions, and personnel in the wake of NCLB and RTTT and the flurry of reforms they have unleashed.²³ Their efforts to reform teacher evaluation offer an excellent example of how SEAs are adapting to the new roles thrust upon them as well as the ways in which ongoing capacity gaps continue to impede their work. Improving teacher quality has become the centerpiece of the Obama administration's education agenda and of the contemporary school reform movement. The past few years have highlighted how difficult this work is and how short timelines and limited staff and funding complicate it further. It is important to recognize that the six early adopter states discussed here are not a random or representative sample of states. By choosing to apply for a RTTT grant, they both self-selected into doing teacher evaluation reform and (because they won) demonstrated a greater initial ability to deliver on it compared with other states. As a result, states that subsequently undertake this work may well struggle even more than these six. But other states can benefit from a close study of the challenges the early adopters encountered in reforming teacher evaluations, and this analysis can inform their efforts going forward.

Dan Weisberg (TNTP) believes that the most important lesson of the past two years has been to prove that new teacher evaluation systems can be implemented successfully. He notes that “there are a number of districts--such as DC and Houston--that have implemented new teacher evaluation systems in a smart, meaningful way and they have really seen results. And by results, I mean they now have systems that really differentiate among teachers based on performance in a relatively accurate, fair way so that they are able to make much smarter, better personnel decisions and they are able to provide teachers with much more helpful feedback on their performance.” Despite these promising examples, however, Weisberg (TNTP) acknowledges that:

What you're seeing is the story in a much larger number of districts where they have not put sufficient time, energy, resources, thought behind implementation. As a result, they haven't seen significant progress from the old evaluation systems that really didn't produce any differentiation—great from good, good to fair, fair from poor—so you still have over 90% plus of teachers being told they are excellent or good in districts that clearly need to improve the quality of education for their students.

In general, states have made a lot of progress in setting up data systems, designing new observational rubrics, and training evaluators. Poda (CCSSO) believes that:

States have made a pretty significant amount of progress in preparing principals and other people to be evaluators. They've spent a lot of time and effort on trying to establish the accuracy and inter-rater reliability among the people who are actually doing the evaluations. Training evaluators is something that states have not paid much attention to in the past so they have put a lot of energy into that.

However, a lot more work remains to be done around incorporating measures of student achievement into evaluations, particularly for teachers in non-tested subjects and grades. Poda (CCSSO) observes

23 . For more on state implementation of No Child Left Behind, see David K. Cohen and Susan L. Moffitt, *The Ordeal of Equality: Did Federal Regulation Fix the Schools?* (Cambridge, MA: Harvard University Press, 2009); Paul Manna, *Collision Course: Federal Education Policy Meets State and Local Realities* (Washington, DC: CQ Press, 2010); and Patrick McGuinn, *No Child Left Behind and the Transformation of Federal Education Policy, 1965–2005* (Lawrence, KS: University Press of Kansas, 2006).

that “states have made progress towards implementing student achievement measures but there are probably only about seven states that are actually doing this full scale right now; the rest of them are still working on it.” States are also struggling with how to adapt professional development to the new evaluation process and with achieving meaningful differentiation in teacher ratings.

Nonetheless, it appears that early adopter states are beginning to settle in with their new evaluation systems. Heyburn believes, for example, that as Tennessee is moving forward with its fourth year of implementing the new evaluation system, teachers and principals are “generally comfortable with the process, and specifically with the rubric that they’re using for the observation component, which is really the part that drives the feedback and the instructional improvement. Observations are taking place and feedback conferences are happening following those observations. We collect and track that data down the indicator level for most of our districts, unless they’re using a third party data system. Additionally, through our work with superintendents, supervisors, principals and teachers, and through the surveys that we do, we can see that these things are happening and folks feel pretty comfortable with the process.” The next—and crucial—step in the development of these evaluation systems will be maximizing the educational benefit of all of this new information. Heyburn (TDOE) noted that:

As we think about our work moving forward, it is really about trying to help people move beyond the process point and really dive more deeply into the quality of feedback and instructional support the teachers should be getting based on the observation and the other data points that make up the evaluation system. In surveys of our lower performing teachers, we see that those teachers in particular are not always getting the quality of feedback that they feel is actionable or the number of observations necessarily that they should be getting.

Several of the interviewees for this paper emphasized the importance of communication and messaging around the new evaluation systems to ensure that the potential benefits are widely understood. Having accurate performance data for teachers has many implications across the human capital continuum from staffing decisions, to differentiated pay plans, to who is promoted into teacher leadership roles. There also needs to be messaging around the fact that the evaluation systems inherently are never going to be perfect and that states are going to need to learn and make changes to the system as they learn over time. NJ’s Shulman emphasized that “there has to be multiple messages, multiple messengers, and multiple vehicles for getting communication down to the classroom level.”

A common refrain from the field is the need to set realistic expectations around the new teacher evaluation systems—both in the sense that people realize that it is complicated, difficult work during which mistakes will be made and also that getting the new systems operating smoothly and effectively will take several years. Anthes (CDE) remarked that:

The last two years have been helping people understand the “what” of the system, what are all the tools, what are all the requirements, what are you expected to do and now we really need to take the next two years to get deep into the “how’s.” So now you know what the requirements are, how do you do it well? And what tools and supports can we provide to help you do it well. And we’re really grappling with how we get deep into practice from the state department level. I think assessment literacy is just a big issue—if we’re going to be basing a lot of these evaluations systems on measures of student learning then we have to help educators be smart about putting those measures of student learning together.

It is also important for SEAs to communicate that these new systems are not primarily intended to be punitive for teachers but rather to improve professional development and classroom instruction. The positive potential of these new evaluations systems is unlikely to be tapped, however, unless evaluators use them to meaningfully differentiate teachers on the quality of their instruction. This is clearly something that states are still struggling with, as Tennessee, Rhode Island, Florida, Indiana, and Michigan all rated more than 95% of their teachers as effective or highly effective during the 2013-2014 school year.²⁴

The 2012 CAP report on teacher evaluation emphasized the need for states to learn and adjust their evaluation systems as challenges emerged during the early years of implementation. Encouragingly, this appears to be happening. NCSL's Exstrom observed that:

I think in general, states are making progress. They're having good conversations about what a good evaluation is and isn't. So many states, just right out of the chute, used the 50% measure, and as the MET study and some other research came out states have gone back to the drawing board a little bit. They're discussing whether or not that is the right mark or whether they can pare that back a little bit, and what other pieces should be included in the performance measures...I don't think that we've come across a state that has said, 'Hey, I think we have the perfect evaluation system.' But what's encouraging to me is just seeing the opening of the conversation around this and really diving more deeply into what a meaningful teacher evaluation looks like.

States are learning from one another as well. New Jersey's Shulman (NJDOE) noted that "we've had the benefit of really learning from some states like Tennessee and Delaware that have been out in front of this work. And I think it's as much about learning what not to do as it is about what to do." It is hoped that sharing these observations from the SEA officials who are on the front lines of the implementation of new teacher evaluation systems will contribute to the continuing refinement and improvement of these important but challenging reforms.

Policy Recommendations

Develop SEA capacity.

Given limited resources, state leaders have to think about how to reallocate existing SEA staff and budgets to focus on new responsibilities, build capacity, and eventually bring work that is funded by external grants on-budget. As they do so, they should consider comparative advantage and economies of scale—where the state can provide something that districts cannot. Providing technical assistance and policy interpretation, creating networks for information sharing, expanding assessment portfolios, and establishing online training modules are several areas where SEAs and SBEs could add real value. States should reorganize their education agencies (as Tennessee and New Jersey have) around discrete functions rather than funding streams and create human capital offices that can integrate the recruitment, training, evaluation, and professional development of teachers. Given the distance-literal and figurative—of SEAs from districts, it is important to create intermediary structures—such as the county offices and regional achievement centers (RACs) in NJ and the Intermediate Units in PA—to provide differentiated and targeted support on a regional basis.

²⁴ Caitlin Emma, "Rating Teachers now As Easy as 1, 2, 3," *Politico*, September 1, 2014.

Provide rigorous and continuous evaluator training and certification.

The RIDE has developed a promising approach to providing ongoing training around the evaluation work. Every summer they run training institutes for all evaluators: a two-day session for veteran principals and a four-day session for people who are new to being a principal or to an evaluation role and need more comprehensive training. It also offers “calibration sessions” during the academic year where they had a team from their office go into a district and work with their leadership team. The sessions focused on setting student learning objectives, observing teachers, providing feedback, and scoring learning objectives. Evaluators should also have to pass a certification test and annual re-certification tests—as they do in Tennessee—to demonstrate their readiness to conduct high-quality observations and ensure inter-rater reliability. In addition to training and certification for evaluators on the front end, it is also important for SEAs and districts to monitor the evaluation results on the back end by looking to see if evaluators are achieving a meaningful distribution of observational scores and how well-aligned those scores are with student achievement data. Tennessee’s SEA analyzes the data to identify schools that have a pattern of ‘misalignment’ and offers them optional support in the form of a TEAM coach from the SEA.

Support principals with their new responsibilities.

As the primary evaluators, school principals are crucial to ultimate success of these new evaluation systems. However, finding enough time in already busy schedules for principals and teachers to do the lengthier and more numerous evaluations, to have the conversations about the results of the observations, and to find ways to use that information to modify and improve instruction is a major challenge. Some states have tried to redefine the principal’s role to reallocate some responsibilities or provide some external capacity to help them. One such example is the Pennsylvania Inspired Leadership (PIL) Program which is a statewide, standards-based continuing professional education program for school and system leaders that focuses more than traditional programs on evaluation skills and using evaluation data to improving instruction. In Rhode Island, some districts have hired a central office staffer to help with the evaluation work and others have created new collective bargaining agreements whereby teacher leaders can help with the observations. Colorado has established a process whereby a non-principal can be trained to become an approved evaluation provider.

Move from evaluation to coaching and instructional improvement.

Once new evaluation systems are operational, SEAs and districts need to ensure that the new information is used to better guide personnel decisions and instructional improvement. In this regard, it is important that the principal evaluation system be aligned with the new teacher evaluation system to ensure that principals appropriately prioritize and are incentivized to assess and coach teachers with rigor and objectivity. In Pennsylvania, for example, their new principal evaluation piece “The Framework for Leadership” followed a year after the introduction of the new teacher evaluation system. Tennessee also redesigned its principal evaluation system to better align with its new teacher evaluation system. For evaluations to be used to inform classroom instruction, teachers need to be provided with differentiated, targeted professional development that can accommodate the wide range of academic disciplines, grade levels, student demographics, and instructional specialists (i.e. ESL and special education teachers). Teachers and principals are being asked to use data—from student assessments and their own evaluations—to create targeted interventions that can drive improvement in student achievement. But they often are not adequately trained to accomplish this task. Creating professional learning communities among groups of educators working in the same subject and/or

grade level can be very helpful, as can providing principals with professional development or coaches to assist them in understanding how to analyze and utilize the new evaluation data.

Create centralized data collection and reporting systems.

A crucial piece of infrastructure for the new evaluation systems is the data collection and reporting systems that districts and states use to gather, analyze, and disseminate the information collection about teacher performance: observations, student surveys, and student growth scores. This is a place where scale is helpful and the state can come up with solutions that are going to be more efficient and reliable than each district dealing with it on its own. Rhode Island, for example, used its RTTT funds to develop the Educator Performance and Support System, a platform that helps districts schedule evaluations, collect data, and provides evaluators with all of their data at an individual and aggregate level. Colorado has also built a management performance platform to help districts manage the data from their evaluations; use of the system is free but optional for districts in the state. In their quest to protect student privacy, however, states have to avoid creating data privacy laws (such as those enacted in Louisiana) that are so restrictive that they could potentially preclude policymakers and administrators from using the student assessment and teacher evaluation data in the ways it was originally intended.

Provide clearinghouse of Student Learning Objectives.

Most teachers work in untested grades or subjects. Figuring out how to measure student achievement or growth in their classrooms remains perhaps the biggest problem confronting the new teacher evaluation systems. SEAs can play a productive role in identifying and designing assessments that are aligned with state learning standards. In Tennessee, for example, the Department of Education developed alternative growth measures that are optional for districts to use, including in world languages, physical education, health, fine arts, special education, pre-K, and kindergarten. States vary widely in the extent to which they have created SLOs and aligned measures and in how centralized the assessment process is. Pennsylvania piloted a voluntary SLO process for districts in 2013–14 that was mandated in 2014–15. The PDE worked with an expert to design training, resources, and templates. Pennsylvania then trained their trainers and piloted the system. The state vetted the models that came from the pilot and provided the exemplars and supporting resources to districts free of charge in 2014–15.

Communicate and engage with stakeholders.

Educators have long complained about the silos in their SEAs and district central offices and their isolation from the field. These concerns underscore the need for effective lines of communication—horizontally and vertically. Given the interconnectedness of teacher evaluations with standards, assessment, and curriculum, SBEs and administrators in SEAs and LEAs must ensure that these different areas are aligned. SEAs also must be accessible to teachers and principals and answer their technical questions promptly. SEAs need to actively engage them in building, piloting, and refining the new evaluation systems. Such engagement will produce a better system and also give stakeholders ownership and buy-in in the system. New Jersey's Evaluation Pilot Advisory Committee and the evaluation advisory committees in each district appear to have been effective in this regard. Operating as they do at the top of the state education governance structure, SBEs have an important role to play in communicating with parents and teachers about what the teacher evaluation changes mean and why they are necessary.

Align teacher evaluation systems with new assessments.

Implementing new teacher evaluation systems is a major undertaking in its own right, but most states and districts are simultaneously rolling out the new academic standards and aligned assessments. This further strains SEA and LEA capacity and emphasizes the need to think carefully about the sequencing of rollouts of new evaluation systems with interconnected reforms. There is a crucial role here for SBEs as they set state policy; it is imperative that core education policies are well-aligned and stable over time. Teachers and administrators in the field can become disillusioned when major policies become disjointed or unexpectedly changed in the middle of being implemented. Tennessee, for example, announced that it would not implement the PARCC assessments at the end of the 2014–15 school year as planned but that it would continue to implement the Common Core State Standards. The state’s standards and assessments are therefore misaligned, and educators believe they cannot be fairly evaluated on the new standards with old tests.²⁵ In a Colorado program, 13 “integration districts” assign teacher leaders as “integration liaisons” to implement new evaluations using a systems thinking approach that integrates new academic standards, assessment, and evaluations. With the help of their SEA, the district leaders met at several professional development gatherings during the year. By piloting the new teacher evaluation systems in advance of “going live” statewide, implementers have been able to identify and resolve problems that emerged and give teachers and principals time to adjust to the new system and their roles within it.

Learn from successes and struggles of other states.

While the design of new teacher evaluation systems varies considerably from state to state, there is much that states can learn from one another as they undertake this work. LEAs, SEAs, and state boards must be forthcoming about what is working and what is not. The Reform Support Network created by the U.S. Department of Education organized valuable convenings of RTTT grantee states where different approaches to teacher evaluation could be shared and discussed. The October 2014 announcement by the Department that it was creating a new Office of State Support is a positive step towards expanding this kind of technical assistance and policy learning to all states.

25 Lauren Camera, “Tennessee Teachers Chafe at Common-Core Uncertainty: Mixed Signals Given after Years of Training,” *Education Week*, December 10, 2014.

Appendix: Interviews Conducted As Part of Research

(Note: While original interviews were conducted in Summer 2014, all interviewees were re-contacted in January 2015 and given the opportunity to revise and update their remarks.)

Katy Anthes

Executive Director of Educator Effectiveness
Colorado Department of Education, July 23, 2014

Shannon Black

Director of Talent Management
Metro Nashville School District, Tennessee, June 23,
2014

Michelle Exstrom

*Education Program Director, Teaching Qualifications
and Effectiveness*
National Conference of State Legislatures, July 28,
2014

Sam Franklin

Executive Director Office of Teacher Effectiveness
Pittsburgh Public Schools (PPS), July 24, 2014

Tim Gaddis

*Assistant Superintendent for Teaching, Learning and
Assessment for Williamson County Schools*
Former Director of Educator Evaluation
Tennessee Department of Education, July 14, 2014

Mike Gradoz

Director of Educator Effectiveness
Colorado Education Initiative, July 22, 2014

Patricia Hardy

Teacher Effectiveness Project Lead Consultant
Pennsylvania Department of Education, July 24, 2014

Sara Heyburn

Assistant Commissioner, Teachers and Leaders
Tennessee Department of Education, July 16, 2014

Sandi Jacobs

Vice-President and Managing Director of State Policy
National Center on Teacher Quality, July 17, 2014

Angela Kirby

*Director, Harrisburg Training and Technical
Assistance Network*
Pennsylvania Department of Education, July 24, 2014

Tim Matheny

*Director of Evaluation, Division of Teacher and
Leader Effectiveness*
New Jersey Department of Education, July 28, 2014

Janice Poda

*Strategic Initiative Director, Education Workforce
Council of Chief State School Officers, July 28, 2014*

Peter Shulman

*Chief Talent Officer/Assistant Commissioner of
Teacher and Leader Effectiveness*
New Jersey Department of Education, July 28, 2014

Mary Ann Snider

*Chief of Educator Excellence and Instructional
Effectiveness*
Rhode Island Department of Education, July 18, 2014

David Volkman

Special Assistant to the Acting Secretary
Pennsylvania Department of Education, July 24, 2014

Dan Weisberg

*Executive Vice-President and General Counsel,
Performance Management*
The New Teacher Project, July 18, 2014