Reform Support Network

Promoting Evaluation Rating Accuracy

Strategic Options for States

June 2013

With the advent of college- and career-ready standards, students must know and be able to do profoundly more than ever before to be prepared for their postsecondary future. With a growing recognition of the importance of teachers in driving student outcomes, policymakers are placing new emphasis on human capital systems that place teachers at the center. In the last two years, leading States have rolled out the newest generation of teacher evaluation systems with high aspirations. The new systems incorporate multiple measures of teacher performance, including observations and student growth and aim to improve student outcomes, help teachers improve their practice and inform career milestone decisions, such as the granting of tenure or compensation increases.

Framers of these new evaluation systems were reacting to the fact that the typical American school district rated 99 percent of its teachers as effective or better, a condition that TNTP labeled the "widget effect" in its 2009 report of the same name. TNTP faulted evaluation systems because they treated all teachers as interchangeable widgets and failed to produce information about individual teacher strengths, weaknesses and effect on student achievement. As they designed the new generation of evaluation systems, the framers sought to reflect the reality of performance in schools, produce a more even distribution of teachers across a performance continuum and therefore give school districts the means to identify teachers in need of support and those they could promote, reward or deploy in new ways to acknowledge their advanced effectiveness.

Early in the implementation of the new generation of evaluation systems, these aspirations have not yet been met. Preliminary data shows persistence of the widget effect despite substantial changes to the design and implementation of evaluations. Evaluation results from States in their first year of implementation indicate that these systems are not producing ratings that help States, school districts, school leaders and teachers better understand the development needs of individual teachers. A group of State and district officials, teachers and principals, and external experts in educator evaluations and strategic communications gathered in the District of Columbia on February 28, 2013 to examine early results from evaluation systems in State A and State B, discuss why these new systems are not creating a more realistic distribution of teachers across evaluation rating categories, and most importantly outline what States can do to address this challenge. This report summarizes the outcomes of that seminar.

Analysis of State Evaluation Rating Data

Experts reviewed early evaluation data from States A and B as case studies representative in many ways of results across the nation. In State A, 99 percent of teachers were rated as effective or higher in School Year (SY) 2011–2012 on the components of the evaluation system that are not related to student growth (State A did not report summative ratings in SY 2011–2012 given that it piloted the evaluation in a subset of schools). In State B, 97 percent of teachers received summative ratings of effective or higher in SY 2011–2012, despite the fact that States expanded the number of rating categories from two to four and incorporated student growth as one measure of performance.

The Reform Support Network, sponsored by the U.S. Department of Education, supports the Race to the Top grantees as they implement reforms in education policy and practice, learn from each other, and build their capacity to sustain these reforms, while sharing these promising practices and lessons learned with other States attempting to implement similarly bold education reform initiatives.

Although the data are still very preliminary, experts noted the following possible trends that could signal problems with evaluation system design (such as observational rubrics, the number of required observations and local discretion to define rating categories), implementation (calibrating observation ratings or explaining the purpose of evaluations), or both:

- **Observation ratings do not align with student outcomes**. In State A, the evaluation data for the student growth component (Component 5) reveal a greater distribution of teachers across levels of effectiveness than the combined results for Components 1-4 (planning and preparation, classroom environment, instruction, and professional responsibilities), for which an evaluator collects evidence during a teacher's classroom observation. The data also show that teachers who *were ineffective at improving student outcomes* — in other words, who received an Unsatisfactory on Component 5 — received a rating of Effective on the other components at virtually the same rate (94 percent) as those *who exceeded expectations at improving student outcomes* (96 percent).

- **The distribution of ratings remains almost unchanged from previous evaluation systems**. State B's previous evaluation system rated at least 99.9 percent of its teachers effective. That figure has dropped by only two percentage points, to 97.2 percent.

The experts identified a list of possible root causes for the misalignment they observed between student outcomes and educator evaluation data, based on their own observations and experiences in school districts and schools in the early stages of evaluation implementation. First, experts noted that principals often lack capacity to manage their human capital. Many do not have the skill or will to assess the performance of teachers accurately, especially those teachers performing at the lowest possible level. Nor do many have the skills and leadership ability to work with teachers to improve performance. Second, experts described the challenge that States have experienced helping school leaders complete high-quality evaluations. States in the early stages of implementation focus largely on compliance — ensuring that observations are complete and evaluations are turned in on time — rather than on quality. Third, experts identified potential issues with system design. States, they suggested, have given too much discretion to school districts in the design and rigor of observation rubrics and the setting of cut scores for various components of the evaluation. Furthermore, States have not yet achieved a balance between positive and negative stakes; under these high-stakes evaluation systems, evaluators have demonstrated a tendency to err on the side of caution. Fourth, experts noted that managing change is difficult and posited that some States and districts have not yet won buy-in for the new evaluation systems from teachers and principals. Finally, experts described problems of political will that allow the stakes for adults to take priority over the stakes for children.

Strategic Options for States

Given this context, the experts identified a set of strategic options in the form of policies, processes and tools for States to employ to address these root problems. The experts convened for the seminar believe these options have the greatest potential to help States pinpoint the differences between ratings across multiple measures and the causes of these variances. These options can be grouped into the following four categories: improving systems to monitor and analyze data, building school leader capacity, improving system design and implementation, and engaging stakeholders.

Improve Systems to Monitor and Analyze Data

Establish a process for monitoring correlation between student outcomes, observations scores and summative ratings.

States need to build performance management systems that continuously assess the relationship between educator observation ratings and student outcomes in both school districts and schools. In doing so, States should review the subcomponents of observation scores to analyze the relationship between key instructional criteria and student performance measures. For example, sometimes evaluators more accurately rate teachers in the core instructional criteria of the rubric but inflate other scores they find to be less critical to teaching, which masks the actual alignment between student performance and ratings pertaining to instruction. Experts suggested establishing an acceptable range of correlation between ratings and student outcomes, collecting and correlating these data points at regular intervals over the course of the year, and identifying and intervening with districts and schools that fall outside of this range. When intervening, States can work with outlier districts and schools to identify the causes of the discrepancies and provide guidance on how to improve the accuracy of educator evaluations. States also can use the data they collect to identify high performers and best practices, and share these successes with others across the State.

Create data dashboard for local educational agencies/schools to facilitate the inquiry process.

To help school districts manage the quality and rigor of evaluations, experts suggested that States generate a data dashboard with analytics that would help school districts analyze their own data and use the information to inform professional development, refine their evaluation systems, or both. Such a template could include data on fidelity of implementation (for example, on-time observations and timely feedback), distribution of ratings and any formative assessment data. As a part of the Reform Support Network (RSN) and the Quality Evaluation Rollout workgroup, eight States are developing a prototype of such a data dashboard, using questions on the performance of their evaluation systems to help inform their choice of data. Sample questions include the following:

- How do different components of the evaluation system (for example, observations, surveys and value-added measures) relate to each other? To what extent does teacher effectiveness differ across the components of State evaluation systems?
- Are schools and districts consistently implementing the new teacher evaluation systems? Are observation ratings correlated to student outcomes more in some schools or districts than in others? Do some districts or schools demonstrate unusual patterns among their evaluation data that suggest inconsistent evaluation policies?
- Do the observational tools used provide meaningful differentiated feedback to teachers, or are all the items too highly correlated with one another to distinguish between different components of teaching?
- To what extent does teacher effectiveness differ across districts and within schools? To what extent do new evaluation systems yield different pictures of effectiveness than the systems they replace?
- At any given time how many teachers have been observed, and how many observations have taken place per teacher?
- Do observation outcomes vary by the type of observer (such as principal, peer, instructional coach or external observer)?

Build School Leader Capacity

Integrate human capital management into principal evaluation.

New evaluation systems and other reforms have permanently changed the role of the principal. Their shift from building managers to instructional leaders requires a new skill set for principals, including accurately assessing observed teacher performance, using data to talk about instruction with teachers and deciding how to deploy professional development resources. Principal evaluations need to reflect this shift in priority. The ability to rate teachers accurately — and to make human capital decisions based on this data — should be significant components of evaluations for school leaders. This sends a message to principals that accurately measuring teacher effectiveness is urgent and important.

Retrain evaluators who aren't accurately assessing teacher performance through observations.

States or school districts should create mechanisms or routines to identify evaluators who are not accurately evaluating teacher performance through observation ratings and retrain them, ideally prior to conducting high-stakes observations. Norming exercises should be a standard component of evaluator training, and evaluators whose observation scores do not accurately calibrate to the master rating should receive more training before they may observe teachers for evaluative purposes. For example, Tennessee is providing coaches to those schools with the biggest gap between observation and student growth ratings to provide them with the support and guidance they need to improve. Next year the State will begin identifying where evaluators are "non-differentiators" and give all teachers the same observation rating across all rubric indicators, so that districts can retrain these individuals.

The experts suggested basing this additional training on practice at the rater's school, rather than on video analysis of teachers that the evaluator will never observe. This would address a concern raised by experts that evaluators tend to rate teachers they know higher than those with whom they are not familiar. Consciously or not, raters make subjective considerations based on factors beyond the evaluation, including their own relationship with the teacher or the teacher's broader contribution to the school community. In short, evaluators need to practice

accurately rating teachers they know. One option is to send impartial observers to conduct co-observations with school leaders. Another option is for States to require evaluators to submit a video of a lesson they observe along with their rating, at which point a trained impartial observer rates the teacher as well. In either scenario, the goal is to identify those school leaders most in need of support and help them improve their accuracy in the context of the classrooms of teachers they will evaluate.

Develop advanced rater certification to continue evaluator skill development.

The first generation of training for evaluators on educator evaluation systems has largely focused on learning the mechanics of the rubric and successfully applying it to a classroom observation. However, the experts noted that even after evaluators have demonstrated the ability to rate teacher effectiveness against a rubric, they are much less effective at giving feedback that actually helps the teacher improve his or her performance. Furthermore, many of the school districts and States that have had evaluation training systems for more than a year are concerned about drift in evaluation scores over time and their ability to keep all ratings normed against a standard of effectiveness. In response to these challenges, States should develop a "next generation" of rater certification systems that not only helps evaluators accurately assess the effectiveness of their teachers but also enables them to provide useful feedback to teachers that ultimately produces better instruction.

Improve System Design and Implementation

Remove laws and regulations that limit best practices.

In some States, there are legal obstacles to more accurate evaluation ratings. Particularly problematic are those laws that restrict multiple observers, require the principal to be the sole evaluator of teachers, or prevent the inclusion of additional measures such as student surveys in evaluation systems. These promising practices have proven effective in other States and school districts at achieving an evaluation system that yields more accurate ratings. For example, Jason Kamras noted with respect to the District of Columbia's master educators (non-school based, content-focused observers) that "the use of external raters was the single most important policy decision we have ever made." The Gates Foundation's Measures of Effective Teaching project confirms that impartial observers are an important component of a fair and reliable evaluation system: "Comparing [impartial observer] scores with scores done by personnel inside the school is the only way to learn whether preconceived notions or personal biases (positive or negative) are driving the scores."[1]

Engage Stakeholders

Build continuous feedback loops to improve evaluation systems.

States and school districts are collecting broad feedback from stakeholders about how their evaluation systems can be improved. But teachers rarely get the opportunity to provide specific feedback on the observations and follow-up conversations evaluators conduct. As one teacher noted, "The evaluation I get from my principal at the end of the year is not helpful. There should be a tool for teachers to give feedback about their principals and how they are helping them improve." Collecting teacher feedback achieves several positive outcomes: 1) giving teachers a sense of ownership of the process; 2) using the data to constantly improve the feedback that teachers receive; and 3) being able to tell a broader audience that teachers play a fair and meaningful role in the process. The RSN has developed an "Educator Engagement Guide" to support States partnering with teachers in this work. In addition, the RSN is building a "Communications Toolkit" — another resource to help States engage and communicate with stakeholders about teacher evaluation.

Conclusion

Early-implementer States face a range of challenges as they try to ensure that their new teacher evaluation systems accurately assess performance and provide clear developmental feedback. In the process, the systems can help school leaders identify teachers who need support and teachers who can be deployed in different ways because of their high levels of effectiveness. States should not lose the hopes and aspirations that drove them to implement these systems in the first place. The options offered by the experts, presented in this report and eventually deployed by States, might help States address challenges presented by new evaluation systems.

[1] "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." (Seattle, Wash.: Bill and Melinda Gates Foundation, January 2012).

Appendix: Evaluation Rating Accuracy Expert Convening

On February 28, 2013, the RSN convened a group of experts in human capital and educator evaluation to engage in the following activities:

- Analyze evaluation rating results from key States, identify patterns and anomalies, and draw informed conclusions relating to the data sets.
- Describe the skills and knowledge that practitioners must have to increase rater accuracy, and discuss how States can effectively train practitioners in such skills and knowledge.
- Develop communication options for States to address the lack of alignment between observation and student growth ratings.
- Develop communication options for States to anticipate the challenges that eventual alignment of observation and student growth ratings might present for States.
- Identify policies, systems and procedures that can help set expectations and create an environment to ensure evaluation results are consistent, reliable and accurate.
- Identify and discuss the data and related systems that schools, school districts and State leaders need to ensure evaluation rating accuracy.

Session participants included experts in strategic communications and educator evaluations as well as teachers, principals, and State and local leaders involved with accountability and human capital from Delaware; Tennessee; New York State; Los Angeles, California; the District of Columbia; Atlanta, Georgia; and New Haven, Connecticut. In advance of the session, participants read the following materials prepared by the RSN: 1) teacher evaluation data for SY 2011–2012 from two RSN States; and 2) an article from *Education Week* on the outcomes of new evaluation systems ("Teachers' Ratings Still High Despite New Measures").

The RSN led a robust discussion on the challenges of and strategies for achieving better alignment between educator evaluations and student outcomes. By the end of the day, the experts developed a set of strategic options for States to pursue to ensure that their educator evaluation systems achieve the goal of improving teacher effectiveness in preparing students for college and the workforce.

Participants

Experts

Stephanie Aberger, Manager, Align TLF Training Platform, *District of Columbia Public Schools*

Tequilla Banks, Executive Director, Department of Teacher Talent and Effectiveness, *Memphis City Schools*

David Guarino, Partner, *Melwood Global*

Kyle Hunsberger, Teacher, *Los Angeles Unified School District*

Jason Kamras, Chief of Human Capital, *District of Columbia Public Schools*

Jatisha Marsh, Teacher, *Atlanta Public Schools*

Amy McIntosh, Senior Fellow, *Regents Research Fund: New York State*

Karla Oakley, Senior Strategist, *TNTP*

David Pinder, Principal, *District of Columbia Public Schools*

Tinell Priddy, Senior Master Educator, IMPACT, *District of Columbia Public Schools*

Christopher Ruszkowski, Chief Officer, Teacher and Leader Effectiveness Unit, *Delaware Department of Education*

Larry Stanton, Consultant, *L. B. Stanton Consulting, Inc.*

Maggie Thomas, Senior Master Educator, IMPACT, *District of Columbia Public Schools*

Glen Worthy, Principal, *New Haven Public Schools*

Reform Support Network

Phil Gonring, Principal

Heidi Guarino, Consultant

Bill Horwath, Consultant

Sarah Johnson, Manager, Teacher and Leader Effectiveness/Standards and Assessment Community of Practice

Kate Sullivan, Policy Analyst

U.S. Department of Education

Marciano Gutierrez, Washington Teaching Ambassador Fellow, *Office of the Secretary of Education*

Brad Jupp, Senior Program Advisor, *Office of the Secretary of Education*

Aaron Pinter-Petrillo, Technical Assistance Team, Implementation and Support Unit, *Office of the Secretary of Education*

Addendum – Evaluation Rating Accuracy State Convening

On April 15, 2013, leaders from four RSN States (States A, B, C and D) that implemented new evaluation systems in the 2011-2012 school year met to analyze their evaluation rating data, identify common challenges and exchange feedback on proposed State action plans designed at the convening to address the challenges. The strategies that States chose to meet these challenges largely mirrored the set of strategic options that emerged from the expert convening in February.

Common Challenges across States

Despite their differences, States reported that they share several challenges, including inadequate systems for evaluation data collection and analysis as well as a lack of skill and will among principals and their supervisors to implement evaluations with rigor.

Data collection and analysis systems are inadequate.

States reported that the evaluation data they received from districts is not consistent in content or does not provide the State with the data it needs to monitor the quality of implementation. In State D, not all school districts reported student learning data to the State. State C and State B did not collect evaluation data by component (for example, observation ratings and student growth ratings), which would give them the means to understand differentiation among teachers at each summative rating level and the relationship between component-level ratings and summative ratings. Furthermore, no States were collecting observation data at regular intervals throughout the year. All of the States had one deadline for the submission of summative ratings: after the end of the school year. This means that when the State intervenes, it is basing its intervention and support on data from the previous school year, which does not include individual component ratings. Without this level of detail, it is difficult for the State to assess the accuracy of the summative ratings. Finally, States find it hard to identify districts that have rating distributions far outside an acceptable norm, because they are not always certain what that norm should be. State C served as an exception: The State reviewed historical student growth data to produce a baseline for how much differentiation it could expect and compared the results of the 2011-2012 school year evaluations to this same distribution.

Principals and their supervisors lack the skill necessary to implement evaluations with rigor.

Participants agreed that States have not yet equipped principals with the skills they need to differentiate levels of effectiveness in observed performance and thereby produce evaluation results that differentiate overall performance. Evaluator credentialing in State A requires evaluators to pass a series of quizzes, none of which requires the candidates to view video of teaching or demonstrate that their ratings of observed performance fall within a specified norm. State B provides technical assistance to help school districts train principals to be effective evaluators, but few districts have taken full advantage of this support.

Principals and their supervisors lack political will.

States reported that principals and their supervisors do not always recognize the value of new evaluation systems. Many focus on checking boxes, rather than on providing useful feedback to teachers. Large numbers of principals treat evaluation as a compliance activity and not as a tool for improving instruction. In State A, principals complete the required paperwork for the evaluation but are not using it to drive the feedback they give to teachers. State participants also suggested that principals are not yet willing to jeopardize long-standing positive relationships by holding teachers accountable to much higher standards. This reluctance to hold educators to high standards is also prevalent among principal supervisors and district leaders. For example, State B gave districts autonomy to set their cut scores for student growth, and many opted to set a low bar for the first year of implementation. As a result, in 17 percent of districts across the State, 100 percent of teachers were Effective or Highly Effective.

State Action Plans

State teams worked together to produce action plans to address these challenges. The plans incorporate many of the strategies that the experts generated at the February convening. States took a targeted approach in developing their action plans, acknowledging that they cannot solve every problem at once, and that certain strategies may address multiple challenges. The State plans prioritized creating data dashboards to monitor and respond to evaluation data, building the skill and will of

principals and their supervisors to implement evaluations with rigor, and using independent observers where possible to lend objectivity and additional data points to teacher evaluations.

Create evaluation data dashboards to improve monitoring.

State leaders described the development of evaluation data dashboards as a high priority for helping them understand the distribution of ratings. States also plan to use the dashboard as a tool to help local educational agencies (LEAs) analyze and respond to the data on an ongoing basis, intervening in school districts when observation ratings are not normally distributed. Building a dashboard requires that States identify the data they want to collect (for example, ratings by individual component) and the times when they want to collect it (for example, at the midpoint and end of the school year). It also requires setting up mechanisms to collect and review the data, which for some States will require a reallocation of resources by the State education agency (SEA).

Build the skill and will of principals and principal supervisors.

In their action plans, States emphasized the need to train principals and their supervisors to analyze their own data. State D plans to regularly convene its superintendents to review data and discuss trends as a way to build executive buy-in to the work. Three States also plan to clarify expectations by disseminating exemplars of strong practice, including examples of effective post-observation feedback and acceptable rating distributions. State A will publish exemplars of principal post-observation commentary, so that principals better understand State expectations of feedback. State C plans to identify school districts that are effectively differentiating teacher performance and hold them up as models. Finally, States recognized a need to set a standard for rating accuracy for principals and their supervisors and hold them accountable for meeting it. State A plans to train its principal managers on how to talk with principals about the outcome of their teacher evaluations. State B plans to meet with leaders in the 17 percent of its school districts where 100 percent of teachers were Effective or Highly Effective to investigate why these districts did not ensure that the new evaluation system produced differentiated levels of effectiveness.

Use independent observers where possible.

At least two State leaders expressed confidence that their evaluators have received adequate training and can apply observation rubrics effectively. However, all States acknowledged that principals have a difficult time issuing objective ratings to teachers they know. To address this, States are considering training independent observers or making better use of staff who can serve in this role. State A has a team of on-demand development coaches that support struggling teachers and principals, and plans to reallocate them to schools where principals might need help accurately assessing performance through observations. These development coaches would help principals calibrate their ratings within an acceptable norm and teach them how to provide teachers with effective feedback. Similarly, State D is considering repurposing a team of professional development providers with content expertise to serve as additional evaluators.