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.

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.

This publication features information from public and private organizations and links to additional information created by those organizations. Inclusion of this information does not constitute an endorsement by the U.S. Department of Education of any products or services offered or views expressed, nor does the Department of Education control its accuracy, relevance, timeliness or completeness.