EVALUATION & SUPPORT Strategies for Success

6 of 6 in the series

Putting It All Together

Using Multiple Measures To Improve Teaching

Throughout this series, we have provided practical advice on implementing high-quality evaluation systems. This includes creating a comprehensive engagement and communications strategy and increasingly meaningful professional learning opportunities for teachers. But states and districts also need to evaluate teachers using multiple measures and combine those measures into a rating or score that accurately predicts future student achievement. States and districts should use data from evaluation systems to establish a culture of continuous improvement, including studying evaluation data and recalibrating weights of measures when those data do not predict student achievement. This way, school systems can make smarter strategic decisions about educator professional learning, talent management, and equitable access to teachers. This brief will explain how states and districts can use evaluation data and will outline how to combine multiple measures into a single composite rating.

As with other evaluation and support policies, school systems must address several challenges related to using data from multiple measures:

- What are the different ways that states and districts can use evaluation data to support teachers and improve the quality of instruction?
- How can states and districts combine ratings from multiple measures—such as classroom observations, student achievement gains and student surveys—into a composite rating?
- When should states and districts use composite ratings, and when should they use data from individual evaluation measures?

This brief answers these questions and explains how school systems can think through the key decision points involved in using evaluation data from multiple measures. The recommendations in this brief will help states and districts use data in ways that improve the accuracy of evaluation systems and make the evaluation process meaningful for teachers and principals.

RECOMMENDATIONS

Decide when and how to use data from multiple measures.

With a system of multiple measures, states and districts will collect more data on teacher performance than ever before. These data can help state, district and school leaders make better decisions, but first they need a vision for how they plan to use evaluation data to support teachers at different stages of their careers (see Talent Management Continuum at left).

TALENT MANAGEMENT CONTINUUM

Prepare | Recruit | Hire & induct | Develop & evaluate | Retain & reward

The talent management continuum shows the different stages of a teacher's career and development. Evaluation data can be used to make strategic decisions at each stage.

Evaluation data can be used in *aggregated* and *disaggregated* forms. *Disaggregated* data refers to the information teachers receive on individual evaluation components or competencies. For example, on most observation frameworks, teachers receive ratings on several different competencies, such as presenting content clearly, using evidence-dependent questioning, or building a positive classroom culture (see Classroom Observations brief for more detail). Similarly, teachers can set multiple learning goals for their students when writing student learning objectives (SLOs). Each observation framework rating or SLO learning goal is a unique data point that is combined, or *aggregated*, to yield a composite evaluation rating for

a teacher. Both aggregated and disaggregated evaluation data can be used to help state, district and school leaders make strategic decisions and support teachers.

Using disaggregated data. Disaggregated data are most useful when helping connect teachers with professional learning opportunities. The Professional Learning and Support brief explained how district and school leaders can provide teachers with individualized growth opportunities based on their evaluation data. Use disaggregated data to identify each teacher's strengths and areas for improvement, and then provide teachers with targeted feedback and professional learning recommendations. District and school leaders should also use evaluation data to assess the quality of professional learning programs and revise or discontinue those that do not improve teaching and learning.

Using aggregated data. Generally, aggregated data are used to inform high-stakes outcomes, since composite ratings provide the most accurate and complete picture of teacher effectiveness. There are four main ways school systems can use aggregated evaluation data:

- **Making talent management decisions.** Some researchers argue that if evaluation and support systems are going to improve practices, they should be tied to outcomes,[1] such as tenure, promotions, bonuses, salary increases and in some cases dismissal. Many teachers, however, believe that evaluation measures should be accurate and that evaluators should be properly trained before evaluation data are used to make talent management decisions. Teachers' concerns can be mitigated by emphasizing the importance of collaboration and continuous improvement so that educators do not feel like they are competing with each other. District of Columbia Public Schools (DCPS) provides pay raises and bonuses to teachers who receive a composite rating of "highly effective" for multiple years and dismisses teachers who are rated "ineffective" after receiving support and professional development.
- **Determining career pathways.** Career pathways vary across school systems, but they commonly entail rewarding high performers with new roles and additional compensation in exchange for taking on certain responsibilities, such as teaching more students and mentoring new or struggling teachers. Composite ratings can help determine how teachers move along their designated career pathway.
- **Ensuring equitable access to effective teachers.** Effective teachers should work with the students who need the most support. Use composite ratings to identify high performers and offer them incentives—such as signing bonuses or opportunities to work with talented school and teacher leaders—to teach in high-needs areas. School leaders should place students who are below grade level in classrooms led by highly effective teachers.
- **Analyzing teacher characteristics.** Finally, districts can use composite ratings to identify the characteristics of high-performing teachers and use that information to recruit and hire new teachers. For example, districts can analyze evaluation data to assess the strength of different teacher preparation programs and funnel resources to the programs that produce the strongest teachers. At a time when recruitment budgets are squeezed, this analysis can help districts determine where and how they allocate their resources.

Lay the groundwork for combining multiple measures.

With a clear vision for using evaluation data, states and districts will be prepared to grapple with the technical challenges associated with combining multiple measures into a composite rating. To lay the foundation for making these critical policy decisions, states and districts should consider these questions:

- **What measures are used to evaluate and support teachers?** Leading research indicates that multiple measures of teacher effectiveness more accurately predict future student achievement.[2] Consistent with this research, many school systems are beginning to use multiple measures—such as value-added estimates, student learning objectives, student surveys and classroom observations—to evaluate teachers. All of these measures should be combined to produce a composite rating, but more is not always better. Too many measures can make it difficult for school systems to analyze teacher evaluation data, and not all measures are high quality or appropriate for evaluation.
- **How are different levels of teacher performance distinguished?** Most teacher evaluation systems fail to make meaningful distinctions between high and low performers. Strong and struggling teachers alike are

rated "satisfactory" and only a handful of teachers are rated "unsatisfactory." Researchers have dubbed this phenomenon "the widget effect," because it treats teachers as indistinguishable.[3] To break through the widget effect, states and districts should classify teacher performance using at least three different categories. The number of performance categories should be determined by each school system's evaluation data. In other words, state and district leaders should look for natural breaks in evaluation data and ensure that performance categories predict future student achievement.

- **What is the relationship between teacher performance and future student achievement?** Highly effective teachers should make the most significant gains in student achievement. If this does not occur, examine the evaluation and support system—the measures used to evaluate teachers, the processes used to generate composite ratings, the alignment of assessments to curriculum, the needs of students, and so forth—and implement the appropriate changes.

Investigate your options: weights or matrix.

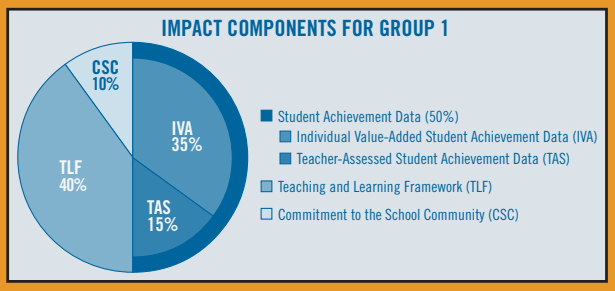
To generate composite ratings for each teacher, states and districts have two main options: using a matrix (or a series of matrices) or assigning each measure a weight. The table below explains how these options work and the tradeoffs states and districts should consider when choosing a method for combining multiple measures.

Some districts use other rules to ensure that their composite ratings are fair and accurate. Denver Public Schools determines composite ratings using a matrix, but principals can exercise their judgment to adjust a teacher's rating. DCPS uses weights, but teachers who do not meet expectations on the Core Professionalism component have points subtracted from their composite rating scores. Regardless of the option chosen, states and districts should review their data to ensure that teachers receiving the highest ratings are making the most significant student achievement gains in the future.

EXAMPLE OF MATRIX OPTION

Instructional Practice and Professional Values	Student Learning Growth 1	2	3	4	5
1	1	1	2	3*	3*
2	1	2	2	3	4*
3	1	2	3	4	5
4	2*	3	4	4	5
5	3*	3*	4	5	5

EXAMPLE OF WEIGHTED OPTION



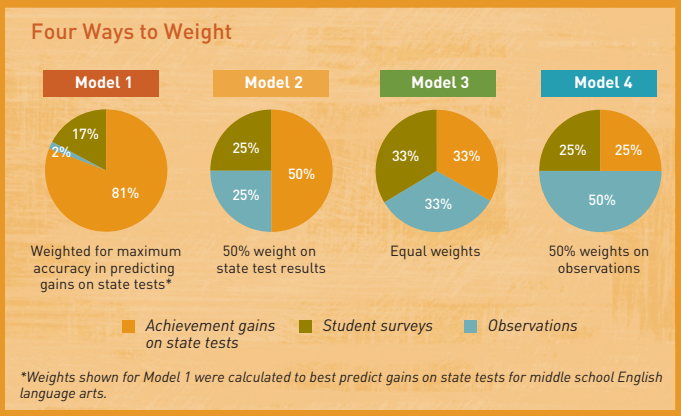
Sources: New Haven Public Schools, available at http://bit.ly/1t2VH5V, and DCPS, available at http://1.usa.gov/VqYeK7.

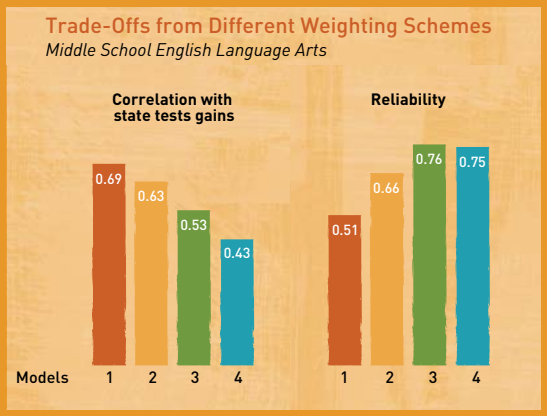
OPTION	DESCRIPTION	EXAMPLES	ADVANTAGES	DISADVANTAGES
MATRIX	A teacher receives a rating on each individual measure. These ratings are plotted on the axes of a matrix. In the example below from New Haven Public Schools, student learning growth is plotted on the horizontal axis, and instructional practice is plotted along the vertical axis. The teacher's composite rating is determined by finding the cell where the ratings converge.	• Rhode Island Department of Education • New Haven Public Schools	• More transparent • Easy to explain to teachers • More uniform and comparable across teachers	• Fewer distinctions between teachers • More difficult to use with multiple measures • Research indicates that matrix approaches introduce more bias into composite ratings[5]
WEIGHTS	A teacher is rated on each individual measure, and each measure is assigned a percent weight. Measure ratings are multiplied by weights and summed together to give an overall score, which corresponds with a performance rating. Weights are also compensatory in that a weakness on one measure may be compensated for by strong performance on other measures.	• Tennessee Department of Education • New York State Department of Education • District of Columbia Public Schools • Hillsborough County Schools	• More precise distinctions between teachers • Accommodates a variety of measures • Research indicates that a weighted approach reduces potential for bias[4]	• Can be less transparent and more difficult to explain and understand • May reduce comparability of teachers in different groups (e.g., teachers in tested versus nontested areas) • Some educators may object to being assigned a number

WEIGHTING DIFFERENT MEASURES

How much should each measure weigh? Generally, measures with uncertain reliability (e.g., parent or family surveys) should have less weight than measures with stronger reliability (e.g., student growth on state tests). The figures below show how weight affects the reliability and predictability of composite ratings. The pie charts on the left show four different ways to weight three measures: achievement gains on state tests, classroom observations and student surveys. Model 1 weights achievement gains the highest (81 percent), while Model 4 weights achievement gains at only 25 percent. Models 2 and 3 weight achievement gains at 50 and 33 percent, respectively.

The bar graphs on the right show what happens to the reliability and predictability of composite ratings when we change the weights of different measures. Model 1 produces composite ratings that predict future student gains but are less reliable; conversely, Model 4 produces composite ratings that are significantly less predictive of student gains but more reliable from year to year. Based on these data, researchers recommend that achievement gains on state tests should weigh between 33 and 50 percent of the composite evaluation rating.





Source: The Bill & Melinda Gates Foundation. (2013). "Ensuring Fair and Reliable Measures of Effective Teaching." Available at http://bit.ly/1nmvbQm.

Establish cut scores.

States and districts that choose the weights option must establish cut scores for each performance category. Cut scores should meet these threshold criteria:

- **Cut scores should be set and adjusted based on empirical data.** When setting cut scores for the first time, districts should use baseline teacher performance data (e.g., from an evaluation pilot).
- **Cut scores should yield ratings that predict future student performance.** Examine the relationship between cut scores and student growth, and be prepared to adjust cut scores if they produce evaluation ratings that do not correlate with future student achievement gains.

If these threshold criteria are satisfied, then the process of setting cut scores can be less technical. States and districts can model several cut score scenarios, and then ask principals and union leaders to reflect on which scenario is the best fit (i.e., the scenario that, in the principal's view, provides the most accurate assessment of teachers in the building).

CONCLUSION

Evaluation systems are giving states and districts access to unprecedented amounts of data on teachers and students. These data can be used to provide teachers with important feedback about their performance and help district and school leaders make strategic decisions about talent management and professional development. Over time, school systems that use evaluation data will better understand their teachers and students and how to help them succeed.

1 TNTP. (2010), "Teacher Evaluation 2.0." Available at http://bit.ly/1gMm883.

2 The Bill & Melinda Gates Foundation. (2013), "Gathering Feedback for Teaching." Available at http://bit.ly/1tzysD8.

3 TNTP. (2009), "The Widget Effect." Available at http://bit.ly/103VkxM.

4 Hansen, Michael, et al. (2013), "Combining Multiple Performance Measures," American Institutes for Research. Available at http://bit.ly/1pO1hGS.

5 *Ibid.*