

**PERFORMANCE
MANAGEMENT** Advantage
Evaluation & Professional Growth

at American Institutes for Research ■

What We Think Matters Most in Evaluating Teachers May Not Be So Important

Surprising Lessons From Redesigning
an Educator Evaluation System



Acknowledgments

The authors wish to acknowledge the efforts of the entire AIR research team involved in this project, led by Lisa Lachlan-Hache. In addition, we benefited from helpful feedback during the preparation of this paper from Ben Backes and Dan Goldhaber.

What We Think Matters Most in Evaluating Teachers May Not Be So Important

Surprising Lessons From Redesigning
an Educator Evaluation System

December 2014

Michael Hansen
Tiffany Chu

Contents

Introduction	1
Goals: Misclassification and Bias	1
Learning From the Pilot Systems	3
Smart Design Principles Reveal Surprising Lessons	5
What Was Thought to Be Most Important But Really Was Not	6
What Was Thought Not to Matter But Really Did.	8
Conclusion	10
References	11
Methodological Appendix	12
Simulated Teacher Performance Measures	12
Coding Candidate Evaluation Systems	12
Misclassification and Bias	13

Introduction

States across the country have been upgrading their educator evaluation systems at a surprisingly rapid pace. Since 2009, a majority of states have made substantive changes to their evaluation systems, at least partially in response to recommendations in the federal Race to the Top initiative¹ or more recently for No Child Left Behind (NCLB) waivers. As a consequence, evaluation systems appear to be one of the most rapidly changing policy issues affecting teachers in recent years.²

In this fluid setting of rapidly evolving evaluation systems, a research team from American Institutes for Research contracted with a state education agency to assist in developing a statewide model for evaluating its educators.³ This paper's authors were among the members of the research team, and we were specifically charged with interpreting the results of a one-year evaluation pilot the state had conducted in a select number of school districts and then making recommendations on how the state should move forward in formally implementing a statewide model.

In the course of this redesign process, we found that differences in the pilot districts' models made large impacts on the distribution of the teacher workforce, independent of performance differences among teachers. As we investigated this further, we learned that the key decision issues for the state ultimately had relatively little bearing on how the evaluation system performed. In addition, procedural details that seemed trivial from the outset proved to have a much more consequential weight than expected. This paper provides a brief synopsis of the process of dissecting the evaluation system. The lessons that emerged from this exercise should be of particular interest to states and districts engaged in designing or modifying their own educator evaluation systems.

Goals: Misclassification and Bias

Before describing the lessons we learned, it is helpful to first understand the motivation and goals behind the state's actions to develop a new evaluation system. Recent research shows that teachers matter a great deal to students and that their effectiveness varies considerably across classrooms and schools. However, typical credentials are not good predictors of an individual teacher's classroom performance.⁴ Hence, process- and outcome-based metrics have taken a prominent role in the U.S. Department of Education's recommendation on strengthening state and local evaluation systems, promoting the use of multiple measures of teacher performance. Yet, exactly how these measures are to be used in a coherent evaluation system is the challenge facing the state's education leaders.

¹ The Great Teachers and Leaders portion of the Race to the Top initiative, which recommended stronger evaluations that considered student performance among other items, accounted for the largest share (28 percent) of the 500 points available to applicant states among all of the selection criteria (see U.S. Department of Education, 2009).

² See documentation of these policy changes over recent years in the report from the National Council on Teacher Quality (2014).

³ For the privacy of the state and the pilot districts involved, participants are kept anonymous in this paper.

⁴ Reviews of the research literature on teacher productivity and its predictors include Hanushek and Rivkin (2010) and Staiger and Rockoff (2010).

Other practitioners similarly engaged in redesigning their own evaluation systems, whether they sit in state or district offices, have by now likely come to the unsettling realization that the federal blueprints for these new evaluation systems are remarkably thin. Beyond a few basic guidelines, there are no established best practices that states or districts can simply adopt in full.⁵ In the spirit of promoting experimentation and homegrown solutions, the state we worked with gave its pilot districts similarly loose reins for the initial pilot year.

The state intended to evaluate these various pilot systems at the end of the year and develop a strategy for moving forward in statewide implementation. It expected to formally endorse (or even prescribe) one or more of the pilot systems, paving the way for districts to use these evaluation scores to differentially manage teachers as they found fit. These strategies may include policies such as targeted professional development for low performers or bonuses for top teachers. Regardless of the exact strategy any district may subsequently choose to pursue, the state's primary goal in recommending a statewide evaluation system is to identify the highest and lowest performing teachers in the district or state in a credible way.⁶

Guaranteeing credibility in the context of educator evaluations, however, is difficult. The implicit assumption behind the use of multiple performance measures is that each measure is informative about elements of quality teaching, but no single metric adequately captures the whole of performance for all educators. Rather, by combining these imperfect measures in some way, teachers can be identified as high or low performers on an aggregated performance measure. In other words, this approach accepts imperfect measures and inferences as inherent characteristics of the system. Yet, the challenge for states and districts upgrading their evaluation systems is to choose an approach that effectively manages these imperfections when categorizing educators' performance.

But what makes an evaluation system effective? Evaluation systems should rely on two guiding principles.⁷ First, the system maximizes accuracy when identifying high or low performers; or, equivalently, it minimizes misclassification between "true" and measured performance. Second, the system minimizes any inherent bias of performance estimates; in other words, performance ratings should not systematically favor teachers at the expense of students or vice versa.⁸

⁵ The Race to the Top guidelines for evaluation systems, which are to include multiple performance measures and categorize teachers into multiple performance ratings (rather than a binary effective-or-not standard), have become the common starting point, but states and districts have departed from this point in myriad ways. Although no best practices in redesigning these systems have been established, the approaches adopted in several frontrunner states and districts have emerged as de facto template models for other states. Some preliminary research evidence has addressed the topic of how to optimally combine these multiple measures (e.g., Mihaly, McCaffrey, Staiger, & Lockwood, 2013), but the science behind making accurate inferences of performance is still evolving.

⁶ The science of combining these measures is not well established; thus, how high or low varies across evaluation systems. Some systems may choose to prioritize student learning gains (or some other objective) first, and others may adopt an approach that weights measures according to the stakeholders' subjective valuation of each metric. A discussion of the trade-offs associated with this choice is beyond the scope of this brief.

⁷ Other features also may be important to states in designing an evaluation system, including transparency or comprehensibility by teachers. As far as selecting a system to perform well on its identification of high- and low-performing teachers, we argue misclassification and bias should be the primary considerations.

⁸ These two concepts are interrelated: A systematic bias will, by construction, increase the level of misclassification. However, even if there were no systematic bias in the evaluation system, misclassification would exist because of the inherent noisiness of the performance measures and may be increased by the evaluation system. See the Methodological Appendix for more details.

What happens if these objectives are not met when redesigning the evaluation system? An inaccurate system fails to reliably differentiate strong and weak teachers and will make little progress on developing teachers' talents and improving the workforce overall. On the other hand, a system biased toward teachers will show a tendency to rate teachers higher than their true performance and could marginalize parents' or the public's confidence in the system to manage workforce quality; likewise, a bias against teachers risks teacher buy-in. Getting both of these objectives right is critical in launching a viable system.

Many design decisions can plausibly affect both the system's accuracy and bias. For example, should raw performance measures be combined so that they can compensate for each other first, before segmenting teachers into the various performance categories? Or, alternatively, what about categorizing a teacher's performance within each measure and then using a decision matrix to determine how the combination of categories corresponds to an overall score?

The state's decision on this and many other points affects the distribution of scores across the teacher workforce; thus, it has the potential to change the system's accuracy and bias in both small and large ways. Yet, how these decisions affect teacher scores can be modeled, and in doing so, states or districts can make informed decisions about which model achieves the system's goals in the most efficient way. The following discussion demonstrates how the state we worked with sifted through the pilot evaluation data to identify the models that would best help its evaluation system achieve success.

Learning From the Pilot Systems

In the study state, a nonrepresentative sample of districts agreed to conduct a pilot of the evaluation system with a sample of willing teachers. As previously noted, the pilot provided participating districts with guidelines similar to those issued by the U.S. Department of Education, with intentionally few requirements. Most importantly for this discussion, the guidelines required districts to collect performance scores on educators from three specific domains, one of which had to be based on student learning in some form. The guidelines also required that each teacher's scores on the three domains be combined in some way to calculate a single categorical summative rating, which takes on one of four categories (1 = ineffective, 2 = marginally effective, 3 = effective, 4 = highly effective).

These guidelines granted the pilot districts considerable autonomy over many aspects of the pilot system's design. Models were designed by district personnel, commonly in consultation with direct input from teachers as well as union representatives in the district. Unsurprisingly, we observed a considerable variety of pilot systems that emerged from this process.

Notable points of variation in the pilot districts' models included the following:

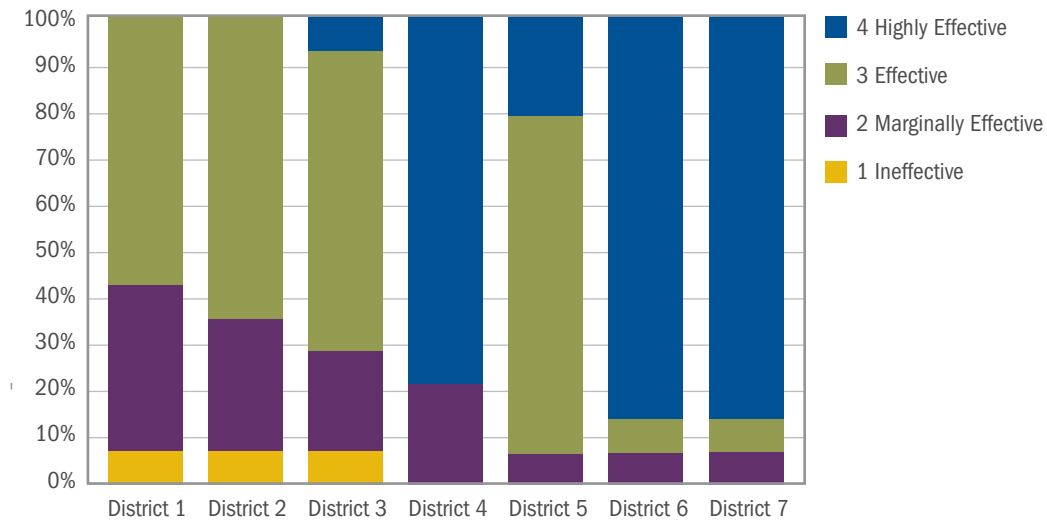
- Different evaluation rubrics to use when observing teachers' in-class performance
- Different numbers of items that were recorded and used to calculate teachers' performance on any given dimension
- Different algorithms that produced a final score within a given dimension of performance
- Different selections of threshold values to combine performance measures into the four category ratings
- Different methods to aggregate scores across performance metrics (e.g., a matrix versus a percentage model)
- Different weights between the different performance measures, either explicitly or implicitly stated, when calculating an overall score

All of these points of difference impacted the classification of individual teachers' performance ratings and the overall distribution of those ratings in some way. In several instances, these differences in the system design that we observed made only slight differences on how teachers were categorized. On other decision points, however, the practical differences between the pilot models were large enough to significantly disrupt the categorization of performance across the entire distribution of teacher scores.

The accumulation of all of these differences on the various aspects of system design meant that **teachers' final summative ratings were almost entirely dependent on the system design, even when the underlying performance metrics were the same.** Figure 1 illustrates the dissonance between the evaluation models in the pilot districts. Each column in the figure represents the distribution of summative ratings across teachers (on the 1–4 scale) that were generated when using each of the pilot districts' models to compute the overall performance rating.

Importantly, we intentionally kept the underlying sample of teachers and their item-level scores on each performance metric constant across all of the columns in this model. Thus, the differences observed here are entirely due to differences in the various districts' evaluation models, and the differences are substantial. For example, District 1's evaluation system in the far-left column provides what appears to be a relatively cynical view of performance: More than 40 percent of teachers are identified as either *ineffective* or *marginally effective*, and no teachers are rated as *highly effective*. Contrast District 1's score distribution with what appears to be an optimistic appraisal in District 7's distribution in the far-right column, where roughly 85 percent of teachers are considered *highly effective*, and no teachers are rated *ineffective*.

Figure 1. Distributions of Teacher Performance Ratings Based on Different Pilot District Models



Smart Design Principles Reveal Surprising Lessons

The state’s approach of devolving authority over the evaluation system to pilot districts seemed reasonable initially; it allowed districts to come up with approaches that, in theory, could be locally optimal although possibly not globally viable. Even so, the level of variation we observed in teachers’ ratings across the models represented an unacceptable risk to the state, compelling the state to reign in some of the autonomies granted to pilot districts.⁹ Instead of the general guidelines issued previously, the state decided to move forward with a standardized statewide model that was chosen because of its relative accuracy and low bias, while still allowing for some local-level variation on a narrow set of dimensions.

The state made a series of design choices to move from a varied collection of pilot models to a viable and effective evaluation system through an iterative process, guided by the objectives of minimizing misclassification and bias.¹⁰ In the midst of this back-and-forth, we were surprised to learn that the design decisions the state (and we) thought would be most critical in optimizing the model turned out to be relatively innocuous. On the other hand, seemingly mundane, off-hand design decisions had disproportionately large influence on the system’s overall misclassification and bias levels. We discuss these surprising realizations in turn.

⁹ Local-level variation in the evaluation system, interacted with the dynamics of a mobile teacher workforce, could also potentially have adverse effects on workforce quality if teachers were to shop for less onerous evaluation systems. Under certain conditions, this could result in a race-to-the-bottom scenario where no district actually discriminates on teacher quality, which would totally undermine the primary goal of the evaluation system.

¹⁰ Sifting through these various candidate models can seem challenging but can be simplified by focusing on a few different metrics on how well the models fit the target distribution. Please see the Methodological Appendix for details of how these simulations were conducted for this paper.

What Was Thought to Be Most Important But Really Was Not

Two issues were particularly important in the state's redesign process from the beginning. Given that these issues are among the more prominent ones we have seen discussed in debates and media coverage on revamping evaluation systems, we expect that most states and districts place a similarly high priority on these areas.

The first issue was how heavily to weight the performance metric that used student learning as its outcome. The state wished to assign more weight to this measure in order to ensure that the new system was sufficiently rigorous and to signal that the system took student learning seriously.¹¹ Conversely, the teachers union representatives pushed to minimize this weight in favor of higher weights on observation and professionalism measures.

The second focal issue was whether to aggregate a teacher's scores across the three performance measures using a matrix model or a weighted index on the raw score values. The matrix model first categorizes performance on each of the three dimensions and then combines them using a predetermined decision matrix.¹² The weighted index simply takes the raw scores from each measure and weights them to calculate an overall score, which is used to categorize a teacher's overall performance rating. The matrix model is commonly perceived to be more user friendly and standards based, which gained favor in teachers' esteem;¹³ however, the weighted-index model tends to be slightly more accurate overall.¹⁴

The state was preparing for these two issues to be the major flashpoints in stakeholder discussions regarding the evaluation system redesign. However, the anxieties of both the state and union representatives were out of proportion to how these two issues actually influenced the overall distribution of teacher performance ratings.

¹¹ Importantly, the state we worked with was redesigning its evaluation system as part of its NCLB waiver application to the U.S. Department of Education; thus, higher weights on the learning-based performance measures were considered in order to increase the likelihood of a favorable review of the evaluation system.

¹² For example, based on performance measures, a teacher may be deemed *effective* in student learning or *highly effective* on in-class observation, and these are the only salient scores for determining the teacher's overall performance rating. Whether the teacher just barely met the standard for *effective* performance or just missed the *highly effective* rating is irrelevant. In other words, performance scores are not directly compensatory across dimensions, as they are when using a weighted index (the percentage model).

¹³ The weighting of measures and the use of a matrix model also are interrelated. One of the features of the matrix model is that it bakes the weighting between performance measures into the decision matrices, and by doing so makes the weights invisible, and thus more palatable, to teachers.

¹⁴ There are some important distinctions between these two approaches, and, in general, the weighted-index model should be preferred as it more closely resembles the statistically optimal method to combine noisy measures (for further discussion, see Hansen, Lemke, and Sorensen, 2013). Yet, the differences between these models can be minimized (and are negligible under some circumstances) if a decision matrix is designed that is consistent with the desired relative weighting between the multiple measures.

To be clear, these decisions do have consequences for the misclassification and bias of the system, but these consequences are relatively minor in the larger context of how these models might vary. To illustrate this point, Figure 2 presents a scatterplot that summarizes the results of a simulation exercise in which we evaluated the misclassification (on the x-axis) and bias (y-axis) for a variety of candidate models the state considered.¹⁵ The optimal method in the center of the crosshairs (labeled “optimal”) is the statistically optimal method derived by using the performance measures’ underlying statistical properties and could be considered the target for this exercise. The remaining dots in the figure represent other actual pilot districts’ models (if labeled with the “Dist” prefix) or other models considered in the redesign process (labeled with a letter).

Figure 2. Misclassification and Bias for Candidate Evaluation Models

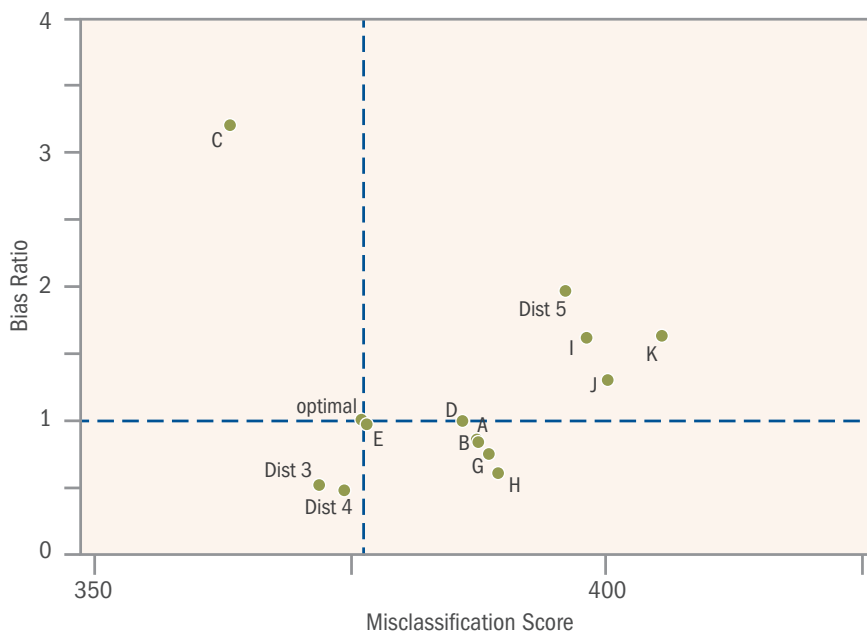


Figure 2 shows one model, labeled E, which is basically indistinguishable from the optimal method. Several other models add either more bias while slightly improving accuracy (Dist 3 and Dist 4) or reduce the accuracy while keeping bias levels roughly constant (A, B, D). Although they vary somewhat on these overall misclassification and bias measures, all six of the models located near the optimal model are sufficiently close to each other that they would perform very similarly in practice. And interestingly, these observations contain both matrix and percentage models,¹⁶ as well as models that vary the weight assigned to student learning based performance measures, ranging from 20 percent to 40 percent. Hence, while the state’s two focal issues did influence

¹⁵ Please see the Methodological Appendix for details on how these measures were created.

¹⁶ Note that the matrix models in this cluster of observations used matrix designs that corresponded to linear weights on the multiple performance measures.

the resulting distribution of teacher ratings, these changes resulted in relatively small marginal differences in the larger picture of all pilot models and those considered during the redesign process.¹⁷

What Was Thought Not to Matter But Really Did

The real issue that made a substantive difference in the system redesign process escaped everyone's notice for a while. After all, rounding numbers to the nearest integer value seems so automatic, why have reason to suspect anything going awry when doing this in an evaluation system?

We found most of the pilot districts chose to categorize performance within a particular dimension (or across dimensions) by taking a simple average of item scores on a particular rubric and then rounding. The problem is that this process inadvertently destroys much of the performance measure's power to discriminate performance. For example, when teachers were observed in the classroom in the pilot districts, they were scored on multiple items that assess an array of practices—presenting a factually accurate lesson, classroom management, and other practices. Each item was scored on a 4-point scale (where 3 equates to effective practice), and in practice most teachers in the pilot districts were given a score of 3 on most items.¹⁸

Simply averaging these scores and then rounding, which many districts did, resulted in all but a handful of teachers being considered effective in this dimension. For example, average scores for classroom observations were commonly distributed on a bell-shaped curve, with a range of values between 2.4 and 3.6. Rounding to the nearest integer only separates out the extreme high and low performers. This means that more than 90 percent of teachers in some districts were identified as effective because their scores fell within the range of 2.5 and 3.49, even though a score of 2.5 signals qualitatively different performance than 3.49.

Replicating this same rounding process across multiple measures and combining them, again as many pilot districts did, implies that virtually all teachers will be considered effective in the end. Revisit Figure 2, and notice the observations labeled I, J, K, and Dist 5 to the upper right of the optimal model. All of these represent models that include

¹⁷ In the interest of full disclosure, we were surprised to learn that some of the largest discrepancies in teacher scores produced by the various pilot models (illustrated in Figure 1) were due to a specific decision matrix used in several of the pilot districts. This particular matrix used an implicit nonlinear weighting scheme in which it was possible for teachers who earned ratings of 3 (effective) on two dimensions at once to be rated a 4 (highly effective) overall. A teacher could earn the overall effective rating only by scoring **less** than effective on at least one performance measure; therefore, the model's nonlinear weights applied a liberal upward bias to the distribution of teachers' ratings. Simply replacing this problematic matrix with one that used an underlying linear weighting scheme to determine the categorical ratings removed much of the observed variability in scores across districts. Hence, although some matrix models are poorly designed, the use of a matrix model alone is not detrimental to the evaluation system if designed to accurately reflect the intended weights across performance measures.

¹⁸ For example, we found that considerably more than 50 percent of the 20+ items used in determining the classroom observation measure for most districts were scored as 3s (effective) across all teachers. Thus, even teachers whose total scores were at the top or bottom of the sample had been considered effective on many, if not most, items.

a variety of weights and combination methods, but the common thread is that all simply average and then round performance scores to the nearest whole number. Relative to the optimal model, this process alone added considerably to the misclassification rate of teachers' performance while placing a liberal upward bias on performance ratings overall.

To counter this tendency to destroy the signal in performance measures, the state instead chose to set thresholds between performance ratings intentionally to meet specific target quantities in each category.¹⁹ Perhaps the state may wish to identify the top 20 percent of teachers as *highly effective*, and it finds the average score at the 80th percentile is 3.25. It may use this as the threshold, rather than the default rounding threshold of 3.5, to distinguish *highly effective* teachers from *effective* teachers. Manipulating threshold values between performance categories is the primary tool the state can use to manage the distribution of performance ratings across all teachers.

Yet, adopting this threshold-setting method raises a thorny issue: Should the state be setting targets for the number of teachers in each performance category? Broaching this issue was treacherous in our experience working with the state, and we speculate this would be true in most states revamping their evaluation systems. The tension arises because the state would, on one hand, like to control the number of teachers in each category for practical purposes.²⁰ Yet, on the other hand, teachers and unions are resistant to the idea that performance expectations are based on relative performance and not based on meeting absolute standards.²¹

In this nexus of conflicting interests, we recommend that new evaluation systems take this threshold-setting approach in the initial years to achieve certain target quantities for each performance rating. The state thus initially controls the outcomes of this system in terms of budget and personnel management. This implies that the scores it takes for a teacher to earn an *effective* or *highly effective* designation, for example, may shift slightly over the first few years as educators and evaluators alike adjust to the new evaluation system and the state fine-tunes its threshold values for performance. If the state desires, after a few initial years of learning, it may then make these new threshold values more permanent—essentially shifting from relative to absolute performance standards.

¹⁹ These thresholds could be set at different levels for different measures, if performance categories are assigned for each metric before combining across the multiple measures.

²⁰ For example, if any state (or district) action is contingent on performance, which is presumably true in most places, the state would like to control the extent of those responses. For example, if a costly and time-intensive professional development intervention is prescribed for teachers rated as *ineffective*, the state would like to limit the number of teachers considered ineffective to limit the state's expected cost for this intervention. Similar logic applies to top-performing teachers (e.g., for retention bonuses). Beyond the practical considerations, the state may wish to target a certain share of teachers in each performance rating for reputational purposes as well. Media reports of upgraded evaluation systems in which almost all teachers are effective or better and very few teachers are rated in the lowest categories have undermined the perception of these systems to actually meaningfully distinguish performance.

²¹ Note that even absolute or standards-based performance measures are implicitly rated on relative scales. For example, classroom evaluators judge teacher performance against expectations developed from what they have observed across other classrooms; student surveys rate teachers on the basis of students' prior history of teachers; objectives for student learning are based on norm-referenced expectations of past performance.

Conclusion

A flurry of recent legislative activity across the United States has put teacher quality into the spotlight, and many states are currently revamping their efforts to promote and evaluate teachers' performance in the classroom. As we aided a state in selecting a statewide evaluation model, we learned that the evaluation system alone in pilot districts made substantial impacts on the distribution of teacher ratings, which was independent of the underlying performance measures for the teachers. To eliminate this variability across districts, we helped the state redesign its educator evaluation system with the overall objective of minimizing misclassification and bias. In this process, we came to some surprising realizations.

First, we found a tendency for both the state and stakeholders to disproportionately focus on elements of the system that had only a relatively small effect on the system's accuracy or bias. Namely, the pilot districts spent most of their energies deliberating about choosing between a matrix model and a weighted-index approach to combining multiple performance measures, in addition to the specific weighting assigned to performance measures based on student learning. As we learned through our simulation exercises, however, the various models being debated performed quite similarly in practice. In general, modest changes in the weights to performance measures make only small differences in the evaluation system's accuracy or bias. And simply choosing a matrix model will not unduly hinder the performance of the system, as long as it corresponds to the intended linear weighting between performance measures.

Second, many in the process were surprised at how important choosing appropriate performance thresholds were, rather than simply using the default tendency to round scores to the nearest whole number. We observed most pilot districts giving virtually all of their teachers *effective* ratings simply because most teachers were considered effective on a majority of items in their rubrics and then rounding these scores. Doing this implicitly destroys the signal in the multiple performance measures and can inadvertently add significant misclassification or bias. Instead, we recommend that states choose threshold values of performance intentionally to correspond to specific segments of the workforce to counter the tendency to score all teachers as effective. These small, but smart, design elements enabled the state to mitigate excess variability in the pilot evaluation systems while also achieving the system's original goals.

Educator evaluation systems are complex, and the science behind them is still catching up to practice. Given the high profile of these new evaluation systems and their potential implications on the strength of the teacher workforce, we hope these lessons help other states and districts navigate this redesign process smoothly.

References

- Hansen, M., Lemke, M., & Sorensen, N. (2013). *Combining multiple performance measures: Do common approaches undermine districts' personnel evaluation systems?* (Report for Value-Added Measurement Services). Washington, DC: American Institutes for Research.
- Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271.
- Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching* (MET Project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf
- National Council on Teacher Quality. (2014). *2013 state teacher policy yearbook: National summary*. Washington, DC: Author. Retrieved from http://www.nctq.org/dmsStage/2013_State_Teacher_Policy_Yearbook_National_Summary_NCTQ_Report
- Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, *24*(3), 97–117.
- U.S. Department of Education. (2009). *Race to the Top program executive summary*. Washington, DC: Author. Retrieved from <https://www2.ed.gov/programs/racetothetop/executive-summary.pdf>

Methodological Appendix

Simulated Teacher Performance Measures

The state provided us with item-level evaluation data from teachers evaluated in select pilot districts, across all performance measures. The correlation matrix between the item-level scores from the districts was used to simulate teacher performance measures across a workforce of 1,000 simulated teachers. Because the correlations between item-level performance measures varied across districts, all candidate evaluation systems were modeled using three different simulated workforces to portray a range of outcomes under various plausible scoring scenarios.

Each of the three simulated workforces corresponded to a different district's performance evaluation data used in the calibration process; the three districts chosen for this were selected because the actual item-level data showed a range in their item-level statistical properties (higher or lower correlations between items, higher or lower variances across teachers, etc.).

Along with the item-level performance measures, a “true” performance measure also was simulated, and it was assumed to have weak, but equal, correlation with all measures of 0.02. This true performance measure was then segmented into four categories based on the value corresponding to the four performance ratings (those up to the 5th percentile were *ineffective* teachers, from the 5th up to the 20th were *marginally effective*, from the 20th to the 80th were *effective* teachers, and those in the 80th percentile or higher were *highly effective* teachers). The performance ratings coming from this true performance measure were considered the target ratings against which teachers' estimated performance ratings were compared in generating the misclassification and bias measures.

Coding Candidate Evaluation Systems

In addition to the item-level performance measures from the selected pilot districts, we also received details on how the performance measures were calculated from the item scores to create a domain-level score, and how these were combined to create overall summative ratings for each teacher. These processes are independent of the actual scoring itself; thus, each pilot district's item-level performance data can be evaluated differently, depending on the specifics of the model being applied to that data. It is the variation in these pilot districts' models, applied to a constant simulated workforce, that results in the various distributions of teacher ratings as depicted in Figure 1.

Beyond the pilot districts' models, other candidate evaluation systems were considered in the process and then coded, which varied specific elements of the evaluation system. For this exercise, three features of the models were focused on and varied to demonstrate the

changes that arise in the misclassification and bias due to changing these parameters. The three features are those discussed in the main text: (a) using a weighted index versus a matrix model to combine the multiple performance measures, (b) the weighting used on the student-learning-based performance measure, and (c) whether districts simply round performance measures to the nearest integer or choose threshold levels corresponding to percentiles as described previously.

Finally, an optimal model also is considered, but for comparison purposes only. This model is statistically optimal in the sense that it minimizes the error in combining the various performance metrics through a least-squares regression approach to determine the weights (methods described in detail in Hansen et al., 2013).

Misclassification and Bias

For each proposed evaluation model during this redesign process, we computed measures of misclassification and bias to score the candidate model. Accuracy (or misclassification) and bias can be measured in a variety of ways in this context, where there are various categories with unequal shares for each category. Given the evaluation system's intended objective to identify high- and low-performing teachers, we used metrics that focused specifically on these teachers. The methods used to calculate these metrics are described.

Misclassification

Misclassification refers to how often a teacher's estimated performance rating fails to capture the true performance rating. Misclassification can arise for several reasons. The primary reason for misclassification is the measures themselves because no measure is a perfect indicator of teacher quality. However, some aspects of an evaluation system (e.g., using simple rounding) may increase the misclassification even further.

The misclassification measure used here is generated by counting the number of categories between the predicted and true performance category ratings among all of those teachers considered either high or low performers (i.e., those truly earning the *ineffective*, *marginally effective*, or *highly effective* rating in the simulated data). For example, a truly highly effective teacher (4 rating) who was identified as effective (3 rating) in the simulation would result in contributing a value of 1 ($4 - 3 = 1$) to the misclassification score.

The optimal model has a total misclassification score of approximately 375 out of 1,000 simulated teachers, 400 of whom are either high or low performing. Thus, high- or low-performing teachers are misclassified by, on average, slightly less than one performance category. This should not be interpreted to mean that almost all high- or low-performing teachers are misclassified because counting the differences in categories implicitly penalizes for larger errors.

Bias

Bias in this context refers to the systematic tendency for the overall estimated distribution of teacher ratings to either overstate or understate performance relative to the true distribution. The true distribution assumes that high and low performers are symmetrically distributed (the top and bottom 20 percent of teachers), but evaluation models can produce a variety of rating distributions that may vary significantly from this. The bias ratio we report here provides a measure of how the estimated distribution varies from the true distribution.

The bias ratio is calculated as the probability that a teacher scores higher than his or her true performance, over the probability that a teacher scores lower than true performance. A score of 1 indicates no systematic bias in a scenario where high and low performers are truly symmetric. Scores greater than 1 imply more effective teachers are estimated relative to the true distribution (overstating true teacher quality); scores less than 1 imply more ineffective teachers are identified (understating true teacher quality).

Note that bias will contribute to the misclassification measure. For example, a systematic upward bias means that fewer low-performing teachers are identified than is true, and, as a consequence, they will be more likely to be misidentified. However, even in cases where there is no bias, misclassification is still present because of the noisy measures and any noise the evaluation system itself adds.

About American Institutes for Research

Established in 1946, American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research on important social issues and delivers technical assistance, both domestically and internationally, in the areas of education, health, and workforce productivity.

About Performance Management Advantage

The AIR Performance Management Advantage, which falls under the larger umbrella of AIR's Educator Talent Management services and resources, is a service line built on a six-step process designed to support a district or state in designing, developing, implementing, and assessing its educator evaluation system.



AIR[®]

AMERICAN INSTITUTES FOR RESEARCH[®]

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
800.356.2735 | 202.403.5000

www.air.org