Carnegie Foundation for the Advancement of Teaching

HOW MIGHT WE USE MULTIPLE MEASURES FOR TEACHER ACCOUNTABILITY?

DOUGLAS N. HARRIS

TULANE UNIVERSITY

CARNEGIE KNOWLEDGE NETWORK

What We Know Series:

Value-Added Methods and Applications

ATIL ADVANCING TEACHING - IMPROVING LEARNING

HIGHLIGHTS

- The most common way to use multiple measures in teacher accountability is through weighted averages of value-added with other gauges of teacher performance. This method has strengths and weaknesses.
- Policymakers should consider a wider range of options for using multiple measures.
- Because the main objective is to accurately classify teacher performance, most discussions of measures of teacher performance focus on validity and reliability. But fairness, simplicity, and cost should also be considered.
- The "matrix" and "screening" methods are somewhat more complex than weighted averages, but they may be more accurate.
- The "screening" method is the least costly and fairest of the three options because it uses value-added measures to improve and streamline other forms of data collection, and it allows final decisions to be made based on the same criteria for all teachers.
- Ultimately, we should assess the method of using multiple measures based on how the options affect student learning, but the evidence does not yet exist to do that.

INTRODUCTION

The idea that multiple measures should be used when evaluating teachers is widely accepted. Multiple measures are important not only because education has multiple goals, but because each measure is an imperfect indicator of any given goal.

For a variety of reasons, states and districts use multiple measures in one particular way: to make personnel decisions about teachers based on a weighted average of the separate measures. Also known as a "composite" or "index," the weighted average provides one bottom-line metric through which teachers can be placed into performance categories. The federal Race to the Top (RTTT) initiative is one reason why states and districts use the weighted average. This competitive grants program required states to hold teachers accountable in a way that made student test scores a "significant factor" in personnel decisions. The meaning of this term is never explained, and the most likely way to meet the vague requirement was to assign large or significant weight—50 percent in some cases—to measures of student achievement growth, such as value-added.

The weighted average approach is also intuitive because people use it in daily life. The Dow Jones Industrial Average combines various stock prices to provide information about the health of the overall stock market; the Weather Channel reports "heat indexes" that combine temperature and humidity to indicate how hot it feels; Consumer Reports measures product quality by combining measures across multiple dimensions; college football rankings are based on an index that combines wins, losses, the quality of opponents, and other factors. These weighted averages allow for simple rankings and comparisons.

Such comparisons are especially useful for some personnel decisions that require an up-or-down vote—we either renew the teacher's contract or not, give tenure or not, promote him to leadership or not. This would seem to require a single measure. To be objective and fair, it would seem that we should place teachers into performance categories based on a combination of performance information, then provide support and professional development to those with the lowest scores and reward those with the highest scores.

While weighted averages are a common and intuitive approach for using multiple measures, there are other options that have their own advantages. In this brief, I also consider the "matrix" and "screening" approaches, which do not involve combining multiple measures.

Unfortunately, unlike in the other briefs in the Carnegie Knowledge Network (CKN) series, there is little evidence about the efficacy of these various approaches. Yet it is still important to show the full range of options and provide a way of thinking about their advantages and disadvantages. This requires breaking out of the narrow measurement perspective. Validity and reliability are worthy priorities, but when we consider methods other than weighted averages, it quickly becomes clear that other criteria—simplicity, fairness, and cost—are also important.

After describing and comparing the weighting, matrix, and screening methods below, I discuss their strengths and weaknesses according to all the above criteria. More than anything else, this brief contributes some new and concrete ways of thinking about how we use value-added and other measures in accountability systems.

WHAT DO WE KNOW ABOUT HOW TO CREATE AND USE MULTIPLE MEASURES?

Weighting approach

The simplest possible weighting for any group of measures is the average of those measures. With an average, each measure is given equal weight: 50-50 for each of two measures, 33-33-33 for each of three measures, and so on. This approach has the advantage of simplicity.[1]

Suppose we have two goals, or elements of effectiveness, for teachers: increasing academic achievement for their own students and contributing to the larger school community so that they help all students. If three-quarters of the definition of effective teaching involves the first goal, then the measure of teacher contributions to academic achievement for their own students should receive a weight of 0.75. The remainder (0.25) would go to the measure of broader contributions. I will call these the "value weights" because they reflect what we value about education and teachers' work, ignoring measurement issues.

Life gets more complicated, though, when we consider that the measures of these two elements of effectiveness vary in their validity and reliability. Suppose we use value-added techniques to measure contributions to student achievement and principal evaluations for contributions to the school community. Also, suppose that principal evaluations do not closely correspond to teachers' actual contributions to the school community, perhaps because the principal judges this based in part on how well the principal gets along with the teacher rather than, as intended, how much the teacher helps other colleagues. In this case, even though contributions to the school community are important, the measure might deserve a relatively small weight because of the questionable validity of the measure.

Value-added measures are also prone to error. In particular, there is growing agreement that random error is the biggest problem—this is mostly what makes value-added measures bounce around so much from year to year.[2] So, even though instructional quality is important, we might reduce the weight on value-added because of this reliability problem.[3]

As a general rule, as I explained in another CKN brief, it is often a waste of resources to collect multiple measures of the same performance construct, except to the extent that additional measures improve validity and reliability when used in combination with other measures or that additional measures are used in part for formative teacher evaluation.[4] This is why a good case can still be made for using both value-added and structured classroom observations. The classroom observations provide more nuanced information about the specific ways in which instruction can be improved (classroom management, quality of feedback to students, etc.), something not possible with value-added. A weighted average of value-added and classroom observations also appears to improve validity and reliability compared with using either measure alone.[5]

Ultimately, if the goal is to accurately assess teacher performance, then we should choose a weight for each measure that minimizes the chances that a teacher will be misclassified, such as a truly low-performing teacher being placed in a middle- or high-performance category. These “optimal weights,” as I will call them, differ from the value weights because each measure varies in its validity and reliability.

Matrix approach

Rather than combining them, multiple measures can also be placed in a “matrix” where personnel decisions depend on the particular combination of measures. In the simplest case, with two measures and two performance categories each, there are four possibilities, illustrated in Figure 1.

Figure 1: Illustration of Matrix Approach

	Performance Measure A	
Performance Measure B	Low A – Low B	High A – Low B
	Low A – High B	High A – High B

For the Low-Low and High-High cells, the performance result will be the same as it was for the weighted average. When you take a weighted average of two low numbers, you have to get another low number. This approach could easily be extended to situations with more than two measures, though this yields more combinations that can be depicted in a simple figure. There could also be other performance categories besides “low” and “high”; Figure 1 is intended just for illustration purposes.

The interesting cases are where we see inconsistencies: the Low-High and High-Low cells. With the weighted average, teachers in both of these categories would be considered average—the lows and highs cancel out.[6] However, as noted earlier, if one of the measures is both more valid and reliable than the other, this makes little sense. Also, it seems highly unlikely that a teacher with high value-added and apparently weak classroom practice is really equally effective as one with low value-added and strong classroom practice.

Screening approach

For a third approach, it is worth looking to the medical profession. It is common for doctors to "screen" [7] for major diseases, using procedures that can identify the vast majority of people who could possibly have the disease. Some patients who test positive will have the disease and some will not—that is, some will be misclassified as "false positives." Those who test positive on the screening test are given another, gold standard test that is more expensive than the initial test but much more accurate. They do not average the screening test together with the gold standard test to create a weighted average. Instead, the two pieces are considered in sequence.[8]

Ineffective teachers could be identified the same way. Value-added measures, like medical screening tests, are relatively inexpensive,[9] but some would argue not very accurate. So, a value-added score should lead us to collect additional information (e.g., more classroom observations, student surveys, portfolios) to identify truly low-performing teachers and to provide feedback to help those teachers improve.

The most obvious problem with this approach is that value-added measures are not designed to capture all potential low-performers.[10] They are statistically "noisy," for example, so many low-performers will get high scores by chance; no additional data would be collected and the low performance would go undetected. Some teachers would slip through the cracks. The false positive rate could also be very large. With value-added as the only screener, the vast majority of teachers would be screened, so that additional information could be collected at some point. Using multiple years of prior data in the screening process would help, but if teacher performance varies over time, then prior years might not be as relevant to assessing current performance. A real trade-off exists here in how to use multiple years of data. For this reason, it would be inadvisable to make value-added the sole screener. Instead, additional measures, such as past performance on other measures, could be used as a screener in conjunction with value-added. If teachers failed on either measure, it would trigger collection of additional information.

There is a second way in which value-added could be used as a screener—not of teachers, but of the classroom observers who rate teacher practice. As with value-added, observations also suffer from validity and reliability issues. Two observers can look at the same classroom and see different things, meaning that "inter-rater reliability" is low. That problem is more likely when the observers vary in how they are prepared for observing teachers or in how they define teacher effectiveness.

The example given earlier is a case in point. The classroom observer might be aware of the teacher's prior performance, and this may color her observations. In general, under traditional evaluation systems, principals give high scores to the vast majority of teachers.[11] Consciously or not, they might think, “I know and like this teacher so I will give her a high observation score.” Or, as the leaders of the schools, principals may worry that low scores reflect poorly on their own performance.

While there is no way to eliminate these types of problems, value-added measures could be used to reduce them. To see how, note that researchers have found consistent, positive correlations between value-added and classroom observations scores. They are far from perfect correlations (mainly because of statistical noise), but they provide a benchmark against which we can compare or validate the scores across individual observers. Inaccurate classroom observation scores would likely show up as being weakly correlated with value-added measures of the same teachers. In particular, if observers fell into the common problem of giving high ratings to almost all teachers, then the comparison with value-added might make this problem evident. Conversely, if observers based their scores on what they already know about teachers’ value-added, the observer ratings would be distorted in ways that make the correlations very high, which might also be a red flag.[12] This approach will work less well when part-time observers are used because they will have fewer observations. A smaller sample size means less confidence in the correlation estimates.

When flags are raised, an additional observer might be used to make sure the information is accurate.[13] In other words, value-added, along with other measures,[14] can help screen the performance of not only teachers, but observers as well. Used in these ways, value-added would be a key part of the process—possibly a “significant” part of the decision according to RTTT—without being the determining factor in personnel decisions.

Evaluating the alternative approaches

The first criterion for evaluating any method of using multiple measures is *accuracy*—whether, for any given definition of teacher performance, teachers are placed in the correct performance categories. This is why so many of the CKN briefs have focused on concerns about validity and reliability; these are what determine the accuracy of performance classifications.

It is also important to recognize that validity is not fixed. When high stakes are attached to measures, Campbell’s Law says the measures will be corrupted (e.g., by changing the way teachers are assigned to students, increased teaching to the test, etc.). It is often hard to foresee how teachers will react and, in this absence of direct evidence, this makes it hard to assess how corruptibility might undermine validity.

Accuracy is only a starting point, however, for understanding how well the use of multiple measures will work in practice. *Simplicity*, while it might encourage manipulation of the measures, is also desirable so that teachers and leaders understand the measures and respond in ways that increase performance. Teachers are also more likely to respond in the hoped-for ways when they believe the accountability system is *fair*.

Educational leaders have also come to learn about the *cost* of teacher evaluations. When expert teachers or principals have to observe teachers, it takes time, a treasured resource in any school. For principals, this might mean less time creating professional development plans or less time with parents. For expert teachers, this might mean less time teaching their own students and, if these really are the best teachers, this is no small sacrifice.

These criteria are somewhat connected. People are less likely to attempt to manipulate the performance system when they see it as fair. On the other hand, it is probably easier to subvert systems that are simplistic and do not incur the costs necessary to minimize corruptibility.

The three approaches for using multiple measures stack up differently on these criteria. The weighted average approach places teachers with different combinations of performance metrics into a single category, while the matrix method has the advantage of being able to handle the inconsistent cases differently.[15] Some of the RTTT winners proposed giving teachers tenure if they are above the bar on *either* value-added or the classroom observation.[16] Alternatively, the rules could preclude placing teachers in the low-performing category if they had high value-added scores, and could not be labeled high-performing if they had a low value-added. The matrix method therefore allows greater nuance, but sacrifices simplicity.[17]

As with the simple matrix approach, the screening idea has the disadvantage of added complexity, but this is offset by lower costs. Schools could achieve similar levels of validity while devoting less time and fewer resources to data collection. The effects on reliability are less clear. Combining measures in a weighted index can increase reliability depending on whether and how the random errors of the various measures are correlated with one another. However, we are really concerned with the validity and reliability of the personnel *decisions*, not the measures themselves, and the screening approach is designed to focus attention on those teachers who are near the margins of each performance category. So, even though screening involves less data collection, it may not sacrifice reliability at all, and may even increase it.[18]

The screening approach also has an advantage over both of the others in its fairness. First, there seems to be an increasing sense among teachers that value-added measures are unfair, so anything that reduces emphasis on them might be seen as more fair. Fairness is also rooted partly in whether the process is applied equally to all teachers. Since most teachers are not in tested grades and subjects, the weighted average and matrix approaches cannot be used for most teachers. But the screening approach can be used in roughly the same way for all teachers, even those in non-tested grades. For teachers who have been in the classroom for more than one year, all the information, including classroom observations, from the prior year could be used as part of the first stage. Even though value-added has some advantages as a screening device, it does not have to be the sole basis for that first stage of the process. And if the second stage is based solely on the classroom observation, for example, then the final performance classification is made based on the same criteria for everyone. In this case, state governments would also be able to worry less about trying to extend standardized testing to grades and subjects for which it might not be appropriate.

Value-added would still play an important role in the screening process, albeit a different and probably smaller one, than it plays now. By moving away from numeric weights, it would be more difficult to show whether value-added is a “significant factor,” but what is more important is whether the evaluation process leads to decisions that are valid, reliable, fair, simple, and inexpensive.

What if each measure captures a different element of teacher effectiveness?

So far, I have written about multiple ways to use multiple measures—based on multiple criteria, no less—but this still over-simplifies matters. I have been implicitly assuming to this point that each measure captures substantially the same elements of effectiveness.

Suppose we defined teacher effectiveness so that it includes exactly two elements: Element A and Element B, each of which has a corresponding Measure A and Measure B. In the earlier discussion, I was assuming that both measures captured the same share of Elements A and B. This overlap might be more or less reasonable when comparing the classroom observation and value-added measures because both are mostly capturing classroom instruction, broadly defined.[19] In the extreme case, where two measures capture exactly the same construct(s), everything I said earlier about the three ways to use multiple methods continues to hold. Making this assumption was a useful starting point, in part because most districts are using measures focused on instruction.

But now take the other extreme. Suppose that Measure A only captures Element A and Measure B only captures Element B. This might be reasonable when value-added is used to measure instruction, while a less structured principal evaluation might capture contributions to the school community that are unrelated to classroom instruction. Student Learning Objectives (SLOs) might also fit this situation if they are intended to capture higher-order learning and if the standardized tests (on which value-added measures are based) capture more basic skills. In these examples, we might say the measures are more “one-dimensional.”

To the extent that the measures capture completely different elements of effectiveness, the weighted average and matrix approaches will make more sense than the screening approach. It would not be sensible to use a measure like value-added in the first stage of a screening approach if the second stage focused on measuring contributions to a completely different effectiveness element such as contributions to the school community. To use the analogy of Consumer Reports, this would be like identifying cars that get good gas mileage in the first stage and collecting information about road handling only if the car got good mileage. In contrast, both the weighting and matrix approaches could be used in these situations to combine separate measures of the separate performance elements.

The way in which we use the matrix approach is also affected by how one-dimensional the measures are. I mentioned earlier that a teacher with high value-added and low classroom observation scores (High-Low) is unlikely to be equally effective as one with low value-added and high classroom observations scores (Low-High). That’s true, but again only to the degree that the measures capture different elements of effectiveness. If the two measures captured each element of effectiveness equally (and if we set aside differences in validity and reliability), then the Low-High and High-Low cases might really be equally effective, and treating those cases the same way might be reasonable.

Whether the measures are overlapping or isolated also affects the quantity of information we have to collect and, therefore, the cost of the system. When the measures overlap, there is less reason to collect multiple measures; they become redundant. But when we need different measures to capture different elements of effectiveness, additional measures become more central, and this drives up costs.

A larger point here is that the choice of the measures and the method of using them are intertwined. The screening approach makes more sense when the measures are overlapping, so the decision about which measures to use cannot be completely separated from the decision about how to use the measures for performance appraisal.

WHAT MORE NEEDS TO BE KNOWN ON THIS ISSUE?

Some progress has been made in recent years in understanding weighted average measures, especially in the well-known Measures of Effective Teaching (MET) project funded by the Gates Foundation.[20] But generally we know very little about how the *use* of the weighting approach really affects teaching and learning, and we know even less about the other methods. To what degree do the constructs of effectiveness being captured by value-added and classroom observations really overlap? Would the complexity of the screening approach be too confusing? These are the types of questions that can only be addressed in states and districts that are willing to alter state rules and in districts that believe alternatives might be more viable.

WHAT CAN'T BE RESOLVED BY EMPIRICAL EVIDENCE ON THIS ISSUE?

The first step in establishing optimal weights—either explicitly in the case of the weighting approach or implicitly in the other methods—is to create the value weights. By definition, value weights are based on what we value rather than the data; therefore, this cannot be resolved with empirical evidence.

HOW, AND UNDER WHAT CIRCUMSTANCES, DOES THIS ISSUE IMPACT THE DECISIONS AND ACTIONS THAT DISTRICTS CAN MAKE ON TEACHER EVALUATION?

The RTTT competition has nudged states and districts into weighted average measures, but policymakers and practitioners may have more, and perhaps better, options than they realize. Theoretically, each of the three approaches laid out here—weighted average, matrix, and screening—has advantages and disadvantages. Empirically, we know relatively little about how well these alternative systems work.

In some sense, these alternatives are not as starkly different as they might seem. They all require multiple measures and require at least an implicit weighting scheme. Also, while I have described the screening approach as more process-oriented, even districts relying on weighted averages have some type of due process procedures in place. There are probably no states or districts where there is not at least some sort of multi-stage appeals process for teacher performance determinations. But the two-stage approach laid out here is still quite different from the appeals processes now in place. A typical appeals process does not use various stages to strategically gather information, and teachers are placed into performance categories in the first-stage, whereas the screening approach would not assign teachers to performance categories until the second stage.

The weighting, matrix, and screening approaches can also be used in tandem. In the description above, I said that “if teachers failed on either measure” then more information would be collected in the screening method. This is essentially a matrix approach for the first stage of the screening process. Also, at the second stage of the screening process, we might still decide to combine two or more measures as a weighted average in making the final personnel decision.

Statistical evidence about validity and reliability is unlikely to get us very far in choosing among these approaches. What we really care about is how educators respond to the measures and that requires a different kind of evidence. To learn how to use multiple measures, we need states and districts to break out of the current narrow weighting mindset and try alternatives that RTTT has so far discouraged. Combined with rigorous evaluations like MET, these studies of the effects of actual implementation on teaching and learning can provide the evidence we really need.

ENDNOTES

[1] Mihaly et al. (2013) also find that equal weighting may be optimal if the goal is to accurately predict future value-added. However, the fact that districts are looking to multiple measures itself suggests that predicting future value-added is not the objective.
See: Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger, and J.R. Lockwood. *A Composite Estimator of Effective Teaching*. (MET Project Research Paper). Seattle, WA: Bill and Melinda Gates Foundation. January 2013.
http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf.

[2] McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4, no. 4 (2009): 572-606.

[3] Of course, if these are the only two measures, we cannot reduce both as the weights have to add up to 1.0. In more complex examples, the correlation among the errors also becomes important. However, with only two measures, the correlation affects the overall validity and reliability of the weighted average, but not the optimal weights. A negative correlation among the errors will improve reliability of the weighted average no matter what the weights are. The situation is much more complex when there are more than two measures.

[4] Harris, Douglas N. Carnegie Knowledge Network, "How Do Value-Added Indicators Compare To Other Measures of Teacher Effectiveness?" Last modified May 2013.
http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Harris.pdf.

[5] If both measures capture somewhat different elements of teacher effectiveness, and both elements are considered important, then the weighted average will improve validity relative to using only one measure. The effect on reliability depends on the correlation in the random errors. See Mihaly et al., 2013, ibid.

[6] This assumes the weights are relatively equal. If almost all the weight is given to Measure A then the teacher's performance on that measure would dominate.

[7] I have written about the screening approach elsewhere.
See: Harris, Douglas N. *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press, 2011.
Harris, Douglas N. "Creating A Valid Process For Using Teacher Value-Added Measures." *Shanker Blog*. November 28, 2012. http://shankerblog.org/?p=7242.
Harris, Douglas N. Harris (2012). "Value-Added As A Screening Device: Part II." *Shanker Blog*. January 29, 2013. http://shankerblog.org/?p=7529.

[8] Since I am using a medical analogy, some might want to call this a "triage" approach. This term fits in some ways but not in others. In both cases, the focus is on allocating resources in cost-effective ways. The higher-performing teachers get less attention just as healthier patients do. On the other hand, there is a difference between this approach and medical triage, as the latter entails devoting few resources to those who are least likely to make it. Instead, part of this point is to collect more information on these struggling teachers so that personnel decisions can be made with confidence and in keeping with legal requirements.

[9] When I say inexpensive, I mean on a per-student or per-teacher basis and for districts and states that take advantage of the larger economies of scale involved in these calculations. To the vendors who provide value-added measures, it costs a similar amount to do it for 100 teachers as it does for 50,000 teachers, whereas with almost any other teacher performance measure, the cost grows in proportion to the number of teachers. (The main cost of value-added measures that is more proportional is having teachers check their student rosters.) This also sets aside the costs of the standard testing regime, which was originally intended for school-level accountability that preceded the movement toward teacher accountability. It also ignores the cost of expanding the testing regime to cover all teachers, partly because I view such an expansion as unwise.

[10] Perhaps the simplest way to identify a screening device is to increase the Type I error rate, so that more teachers are automatically identified as potentially ineffective. But taking that approach to the extreme defeats the purpose of the screening approach: identifying potentially ineffective teachers efficiently. In

addition to increasing the Type I error rate, the usefulness of a screener can be improved by using more prior information about teacher performance, which is also inexpensive for the fact that it has already been collected for prior evaluations.

[11] Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. Brooklyn, NY: The New Teacher Project, 2009. http://widgeteffect.org/downloads/TheWidgetEffect.pdf.

[12] Setting aside the reliability of the correlations, this approach will capture most validity problems. If an observer is an easy or harsh rater (showing up in very low or high average scores) and the observer scores hit the floor or ceiling of the observer scale, then it would show up as a low correlation and raise a red flag. If only one of these conditions holds, then the approach might not work, but these are likely to be the less extreme cases (e.g., an easy rater will only not hit the ceiling when their biases are small). One circumstance in which the correlations might be of little value is when the observer makes different types of errors with different teachers that cancel out and yield a correlation in the expected range.

[13] The comparison between observations is most useful when the two observers are scoring the same instance of classroom instruction, rather than following up on a different day or a different class or subject on the same day. However, having two observers at the same time might pose coordination problems; another way to address this would be through video-taped instruction.

[14] The use of value-added in this way is not meant to preclude or replace other methods for improving the validity of classroom observations (or other measures such as student learning objectives). For example, it is also advisable, albeit costly, to have multiple observers and do calibrations by having observers see the same instances of classroom instruction.

[15] Another dimension sometimes used in these matrices is the level of student test scores. One reason for doing this is that ceiling effects of tests mean that teacher value-added may be less valid for teachers whose students are already at high levels. In that case, the matrix approach allows teachers with high score levels and medium or low value-added to be treated differently. Louisiana is one state that is adopting this approach.

[16] Mihaly et al. (2013) call this the “disjunctive approach.” Alternatively, when high performance is required on both measures, they call it the “conjunctive approach.”

[17] It would seem that one additional advantage of the matrix approach is avoiding the difficult task of creating weights, but this is a bit of an illusion. If teachers in the Low-High and High-Low categories are treated equally, then, implicitly, the two measures are equally weighted (50-50). The matrix approach is also just as costly as the weighted average method because the information still has to be collected on all teachers in order to place each of them into the correct cell.

[18] Reliability could increase if value-added were used successfully to reduce the error in the classroom observations.

[19] When I say classroom instruction, I mean essentially everything happening in the classroom from literal instruction to classroom management.

[20] Mihaly et al., 2013, ibid.

AUTHOR

Douglas N. Harris is Associate Professor of Economics and University Endowed Chair in Public Education at Tulane University. His research explores how students' educational outcomes are influenced by school choice, standards, teacher evaluation, test-based accountability, college financial aid, and college access programs. A former school board member, his research marries theory and rigorous research with the practical realities of schooling with publications ranging from the general interest journal *Science* to the *Journal of Public Economics*, *Journal of Policy Analysis and Management*, and others. His recent book, *Value-Added Measures in Education* (Harvard Education Press, 2011) was nominated for the national Grawemeyer Award. *Washington Monthly* magazine has used his research on college performance measures in its college ratings and David Brooks has cited related work in his *New York Times* column. He has advised eight state departments of education, elected officials at all levels of government, and groups such as the National Academy of Sciences, National Council of State Legislatures, National Governors Association, and National School Boards Association. His work is frequently cited in the national media, including CNN, *Education Week*, *The New York Times*, and *The Washington Post*.

ABOUT THE CARNEGIE KNOWLEDGE NETWORK

The Carnegie Foundation for the Advancement of Teaching has launched the Carnegie Knowledge Network, a resource that will provide impartial, authoritative, relevant, digestible, and current syntheses of the technical literature on value-added for K-12 teacher evaluation system designers. The Carnegie Knowledge Network integrates both technical knowledge and operational knowledge of teacher evaluation systems. The Foundation has brought together a distinguished group of researchers to form the *Carnegie Panel on Assessing Teaching to Improve Learning* to identify what is and is not known on the critical technical issues involved in measuring teaching effectiveness. Daniel Goldhaber, Douglas Harris, Susanna Loeb, Daniel McCaffrey, and Stephen Raudenbush have been selected to join the Carnegie Panel based on their demonstrated technical expertise in this area, their thoughtful stance toward the use of value-added methodologies, and their impartiality toward particular modeling strategies. The Carnegie Panel engaged a User Panel composed of K-12 field leaders directly involved in developing and implementing teacher evaluation systems, to assure relevance to their needs and accessibility for their use. This is the first set of knowledge briefs in a series of Carnegie Knowledge Network releases. Learn more at **carnegieknowledgenetwork.org**.

CITATION

Doug N. Harris. Carnegie Knowledge Network, "How might we use multiple measures for teacher accountability?" Last modified October 2013.
http://www.carnegieknowledgenetwork.org/briefs/multiple_measures/

Carnegie Foundation
for the Advancement of Teaching

Carnegie Foundation for the Advancement of Teaching
51 Vista Lane
Stanford, California 94305
650-566-5100

Carnegie Foundation for the Advancement of Teaching seeks to vitalize more productive research and development in education. We bring scholars, practitioners, innovators, designers, and developers together to solve practical problems of schooling that diminish our nation's ability to educate all students well. We are committed to developing networks of ideas, expertise, and action aimed at improving teaching and learning and strengthening the institutions in which this occurs. Our core belief is that much more can be accomplished together than even the best of us can accomplish alone.
www.carnegiefoundation.org

We invite you to explore our website, where you will find resources relevant to our programs and publications as well as current information about our Board of Directors, funders, and staff.

Knowledge Brief 11
October 2013

carnegieknowledgenetwork.org

Funded through a cooperative agreement with the Institute of Education Sciences. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.