AERA Open
January-March 2017, Vol. 3, No. 1, pp. 1–16
DOI: 10.1177/2332858416685550

Lower Performance Evaluation Practice Ratings for Teachers of Disadvantaged Students: Bias or Reflection of Reality?

Anthony Milanowski
Westat

Value-added estimates of teachers' contributions to student achievement have been criticized for bias relating to the sorting of students to classrooms. More recently, research has raised the possibility that sorting leads to differences in practice evaluation ratings between teachers of more or less disadvantaged and/or higher- and lower-achieving students. Adjusting observation ratings for the relevant characteristics of teachers' classrooms has been proposed as a remedy, analogous to how value-added teacher effectiveness estimates are developed. However, the appropriateness of adjustment depends on the cause of observed differences in average ratings and the use of the ratings. Potential causes include rater bias rubric deficiency, differences in teacher skills and beliefs, and true differences in difficulty of teaching groups of students. The appropriateness of adjustment given these causes and typical uses of the ratings is discussed as well as research needed to identify the influence of the causes.

Keywords: *rater bias, teacher evaluation*

VALUE-ADDED estimates of teachers' contributions to student achievement have been criticized for bias resulting from sorting of students to classrooms. Critics have argued that typical value-added models do not sufficiently control for disadvantages some teachers might suffer if assigned more-difficult-to-teach or lower-ability students when these estimates are used for teacher evaluation (e.g., Berliner, 2014). Perhaps in response, proponents of value-added measurement have pointed out that measures of teaching practice used in evaluation, typically based on classroom observations, may be subject to similar types of bias (J. Cohen & Goldhaber, 2016). The potential for bias could even be greater, because unlike value-added methods, which by definition control for students' prior achievement and typically control for factors such as student poverty and ethnicity, practice assessment ratings are typically used without any such adjustment.

Recent research has found that teachers with disproportionate numbers of some types of students do receive lower practice ratings. Borman and Kimball (2005) found that classrooms with higher concentrations of poor, minority, and low-achieving students were more likely to be taught by teachers with lower evaluation scores. Chaplin, Gill, Thompkins, and Miller (2014) reported negative correlations between scores based on observations of practice and classroom proportions of minority and free lunch–eligible students in a district using observation ratings in a multimeasure performance evaluation system. Lazarev and Newman (2015), using data from the Measuring Effective Teaching (MET) Project, found positive correlations between practice scores based on videos of classrooms and classroom average pretest scores. Steinberg and Garret (2016) have recently taken this line of research forward by examining the relationship between evaluation ratings and students' prior achievement, again using MET data, controlling for teacher fixed effects, which should presumably capture teachers' time-invariant instructional skill. They found evidence that the incoming academic performance of teachers' students influences their performance ratings. Whitehurst, Chingos, and Lindquist (2014), after finding that teacher rankings based on observation scores were associated with students' level of prior achievement, concluded,

> This represents a substantively large divergence from what might be expected from a "fair" system in which teacher ratings would be independent of the incoming quality of their students. (p. 17)

These findings have led to suggestions that such ratings be adjusted for student poverty or prior achievement, analogous to how value-added estimates are typically estimated with controls for preexisting learning and student demographic characteristics (Whitehurst, 2015).

However, there are some reasons to be cautious about adjustment. Despite their findings, Lazarev and Newman (2015) cautioned that

> if less proficient teachers are assigned to classes made up of lower-performing students or if schools serving low-income communities are less successful in retaining effective teachers, then such an adjustment would undermine the validity of an evaluation system by obscuring the real differences among teachers. (p. 2)

Adjustment would also make it more difficult to assess whether disadvantaged students have less access to good teaching. Further, the use of evaluation ratings to improve individual teachers' practice would be undermined by removing the direct link between the rating and the descriptions of teaching in the rubric. An adjusted rating provides less useful information about the level of practice to the teacher or a mentor or coach. Unlike value-added estimates, rubric ratings are intended to be criterion referenced, and adjustment weakens, if not breaks, the reference.

The potential disadvantages of adjustment suggest that it is important to consider why ratings differ. Differences may reflect true differences in teaching. As the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014) caution,

> Subgroup mean differences do not in and of themselves indicate a lack of fairness, but such differences should trigger follow-up studies to identify potential causes of such differences. (p. 65)

This article argues that a decision to adjust ratings to remove effects of classroom composition is premature without a more in-depth consideration of why they differ and the uses to which they will be put. To motivate this consideration, the article first reviews the nature and typical purposes of teacher practice ratings, then describes five potential causes of rating differences and considers evidence for their plausibility. The appropriateness of adjustment for classroom composition is then considered for various combinations of potential causes of ratings differences and potential uses of evaluation ratings. Research that could be used to better understand differences in ratings associated with classroom composition is then discussed.

How Observational Practice Ratings Are Made and What They Are Intended to Measure

Observational practice ratings used for teacher evaluation are typically made on several dimensions of practice (e.g., engaging students, lesson structure and pacing, managing classroom procedures) using 4- or 5-point rating scales (generally called *rubrics*), with each point anchored by a description or set of examples of the behaviors that merit rating at that point. Observers are expected to watch and listen to teacher (and sometimes student) behavior, typically during a fixed number of occasions; record or encode that behavior using notes or checklists; connect the recorded or encoded behavior (often termed *evidence*) to the appropriate performance dimension; and decide which rubric level best fits the evidence collected. In some systems, ratings are made for each occasion of observation (typically a class period), and an overall rating is calculated from them via an algorithm (e.g., averaging). In others, evidence (recorded in notes) is accumulated over multiple occasions of observation, and a rating is given based on the preponderance of evidence. In some systems, ratings are required to be based only on behavior observed during specified periods, sometimes termed *formal observations*. Dimension ratings are then combined to yield an overall evaluation rating.

Ratings have multiple uses. Formative uses include providing feedback to teachers to help them improve their practice or to identify which teachers should receive additional training or professional development. Summative uses include deciding which teachers should be retained, dismissed, or rewarded. Ratings and the underlying rubrics also function to define and communicate a standard of practice and hold teachers accountable for meeting it. Although the different uses made of ratings have implications for judgments about accuracy, bias, and unfairness (discussed later), they all assume that the ratings accurately represent the behavior of a teacher during the occasions of observation. Although ratings cannot completely represent what occurred in the classroom, the observable behaviors are assumed to be the basis for the rating, filtered through the observer's efforts to observe and record, encode, or recollect the behavior; identify the recorded or recollected behavior relevant to each performance dimension in the rating scales or rubrics; and choose the performance level that best fits the observed behavior.

It is important to recognize that there are many potential challenges in measuring teacher performance. Observers may lack sufficient knowledge of the content area to understand how to apply the rubric language (Hill & Grossman, 2013). Teacher behavior also varies across time (Curby et al., 2011; Rogosa, Floden, & Willett, 1984) and even within lessons (Malmberg, Hagger, Burn, Mutton, & Colls, 2010), so that the timing of observation can influence ratings. Observation alone may not provide sufficient information to understand the context or intent of a teachers' practice (Stodolsky, 1984), and some aspects (e.g., ability to engage students) could be better assessed by surveying students (Kunter & Baumert, 2006). Thus, practice ratings based on observation may not completely represent teacher performance.

Potential Causes of Relationships Between Teacher Practice Ratings and Student Disadvantage[1]

Figure 1 provides a framework for thinking about factors that could influence the correlation between classroom composition and practice ratings such that teachers of disadvantaged students on average receive lower ratings.

Three of the factors—student disadvantage; teacher skills, abilities, and beliefs; and teaching conditions—could all have direct impacts on teacher behavior, the intended evidentiary basis for the rating. The others—rater bias and

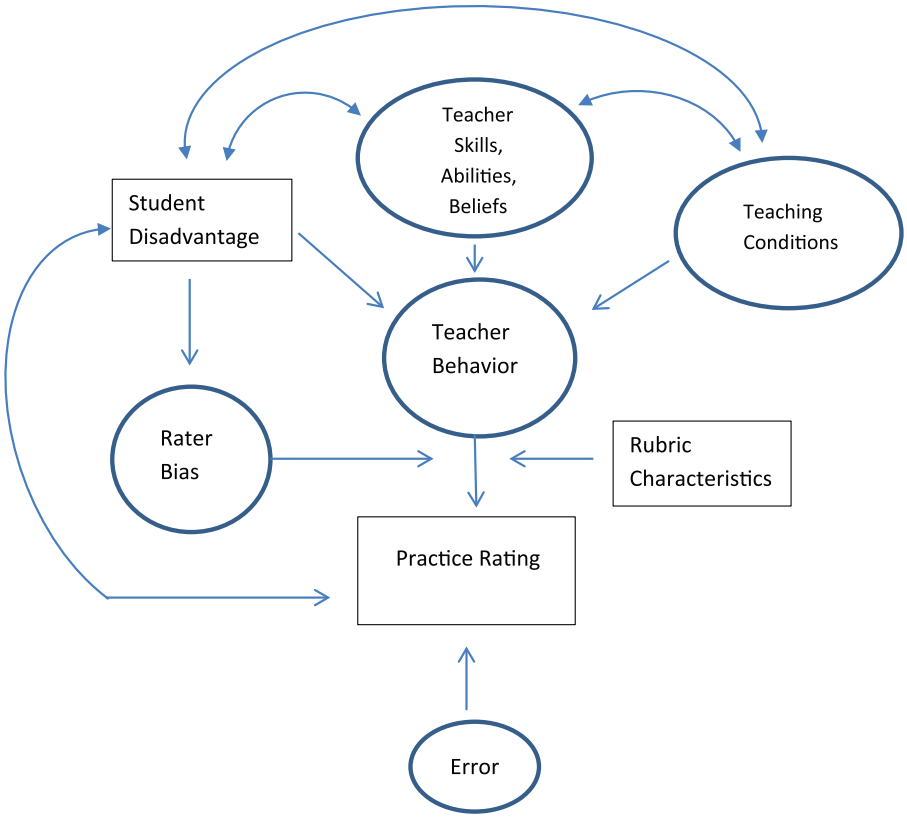

FIGURE 1. *Potential causes of differences in practice ratings between teachers of more or less disadvantaged students.*

rubric deficiency—do not directly influence teacher behavior but rather affect the translation of observed behavior into performance ratings.

Teacher Skills, Abilities, and Beliefs

If teachers of disadvantaged/low-prior-achievement students have lower levels of teaching skill, one would expect that at least some of the relationship between evaluation ratings and classroom composition would be due to this skill gap. These teachers would be less able to teach in ways described at the higher levels of the rubrics. There is a substantial amount of research that suggests that at least some of the relationship between evaluation ratings and student disadvantage is due to lower levels of skill of the teachers assigned these students. There is considerable evidence that poorer students are more likely to have less experienced teachers (e.g., Allensworth, Ponisciak, & Mazzeo, 2009; Clotfelter, Ladd, & Vigdor 2005; DeAngelis, Presley, & White, 2005) and that less experienced teachers are less effective, as shown by lower levels of value added (e.g., Chingos & Peterson, 2011; Rice, 2013). There is also evidence that in some districts these students are taught by less effective teachers (Max & Glazerman, 2014). Less experienced teachers often also get lower practice ratings (e.g., Harris & Sass, 2009; Jacob & Walsh, 2011; Milanowski, Kimball, & White, 2004), although like the relationship between value added and experience, this relationship is strongest in the early years of a teacher's career. The lower experience (and thus lower skill) of teachers of disadvantaged students could thus be a reason for the difference in evaluation ratings. As turnover research (e.g., Boyd, Lankford, Loeb, Ronfeldt, & Wyckoff, 2010; Hanushek, Kain, & Rivkin, 2001) suggests, many teachers prefer teaching less disadvantaged students. Schools with such students would thus receive more applications for vacancies and could be more selective in hiring, leading to differences in teacher skills being correlated with student disadvantage. Steinberg and Garrett's (2016) finding that English language arts teachers with apparently lower levels of skill were assigned students with lower prior achievement supports the contention that at least some of the ratings differential is due to skill differences among teachers.

Teacher beliefs about students may also influence their behavior. Teachers assigned to classrooms with disproportionate numbers of poor students or students with lower prior achievement might also believe that these students lack the prerequisite preparation to allow use of the pedagogy described in the higher levels of the rubric. Or they may believe that these techniques will not be effective for

lower-performing students. Based on these beliefs, these teachers may decide not to attempt to use techniques that are represented at the higher levels of the evaluation rubrics. Note the distinction here between teacher beliefs and the actual difficulty of using these techniques. This distinction is important because reformers have argued that students at all levels can be taught in more constructivist ways and that a primary barrier to this is low teacher expectations. Although this could be seen as a variation of differential teaching skills, it is worthwhile to distinguish low expectations because of the research (e.g., Figlio, 2007; Sorhagen, 2013) and advocacy (e.g., Marchitello & Wilhelm, 2014) around the contribution of teacher expectations to student learning

There is evidence that teachers may not believe that the types of pedagogy described in higher levels of rubrics, like the Framework for Teaching (Danielson, 2007), are effective with lower-performing students. Zohar, Degani, and Vaaknin (2001) found that many teachers believed that teaching activities based on higher-level thinking skills were ineffective for low-achieving students. Warburton and Torff (2005) found that teachers deemed it appropriate that less advantaged students receive fewer high-critical-thinking activities than more advantaged students. Wigfield and Eccles (2000) reported that teachers who believe certain children are incapable of learning are less likely to provide them stimulating tasks that improve their learning. Raudenbush, Rowan, and Cheong (1992) found that teachers believe they are more effective when working with higher-ability students. More recently, Gersheson, Holt, and Papageorge (2016) found that non-Black teachers of Black students have significantly lower expectations for future educational attainment among these students than do Black teachers, which could influence the methods used to teach these students. If teachers have lower expectations for lower-achieving students, or find them less rewarding to reach, it is possible that teachers and students will be less engaged with each other, leading to lower ratings on evaluation dimensions that involve engagement. Steinberg and Garrett (2016) found that ratings on performance dimensions they interpret as requiring substantial collaboration between teachers and students more consistently showed larger differences because higher-achieving students are more likely to be engaged in the learning process.

Teaching Conditions

The general literature on performance evaluation (e.g., Bacharach & Bamberger, 1995; Peters & O'Connor, 1980) has recognized that performance and, in turn, performance ratings are affected by workplace conditions and resources. There is evidence that one workplace condition, class size, affects several aspects of teaching practice, including the type of classroom discussion (Rice, 1999), interaction with individual students and classroom management (Blatchford & Mortimore, 1994), and questioning techniques (Bourke, 1986), although these effects are not always found (e.g., Pong & Pallas, 2001), perhaps because some teachers may not have the skills to make use of the opportunities smaller classes provide (Graue, Hatch, Rao, & Oen, 2007). If class sizes are larger for disadvantaged students, this could contribute to the correlation between student disadvantage and ratings. D. Cohen, Raudenbush, and Ball (2003) argued that instructional materials and resources have an important impact on instructional practice, and there is also some evidence that some of these resources may be less available in classrooms containing more poor and lower-performing students, at least in science (Smith, Nelson, Trygstad, & Banilower, 2013). If disadvantaged students are more likely to be taught in schools or classrooms where teachers have less access to materials or resources that facilitate teaching described at higher rubric levels (due, possibly, to lower funding levels), teachers of these students could tend to receive lower ratings.[2]

Student Disadvantage

In addition to being correlated with teaching conditions and influencing or being correlated with teacher skills, abilities, and beliefs, classrooms with a higher proportion of disadvantaged students could be harder to teach. There is evidence that poor students have often have emotional, social, and cognitive challenges that they must overcome to learn as quickly as more advantaged peers (e.g., Kolb & Gibb, 2015; Lipina & Colombo, 2009) or that affect their behavior in ways that could make then harder to teach (e.g., Yoshikawa, Aber, and Beardslee, 2012). Value-added results appear to confirm that poor students are less likely to learn: Few if any value-added models estimate a positive coefficient for the poverty indicator. As to whether poverty or lower prior achievement makes students more difficult to teach, the evidence seems mixed. Some studies suggest that classrooms with a higher percentage of poor children are less productive (e.g., Ballou, Sanders, & Wright, 2004; Newton, Darling-Hammond, Haertel, & Thomas, 2010). However, Sass, Hannaway, Xu, Figlio, and Feng (2010) reported mixed results for poverty at the school level and no differences in teacher value added when switching between low- and high-poverty schools. Xu, Özek, and Corritore (2012) found that switching between schools with substantially different performance or poverty levels does not reduce teacher performance. Xu, Özek, and Hansen (2015) found that improvements in the effectiveness of new teachers was related more to initial effectiveness than school poverty, suggesting that it may not be harder to learn to teach effectively in high-poverty schools. Fox (2016) found minimal differential effectiveness within teachers by student ability or free lunch status, suggesting teachers may not find it more difficult to teach effectively with poor or lower-achieving students. There is also some evidence that classroom student

achievement growth as measured by value added can actually be greater for students with lower prior achievement (e.g., Protik, Walsh, Resch, Isenberg, & Kopa, 2013), but it is unclear whether this is due to test ceilings or similar artifacts (Resch & Isenberg, 2014).

Evidence about whether it is more difficult to teach as valued by practice rubrics is limited. Polikoff (2015) looked at the relationship of changes in total practice evaluation scores across years and changes in classroom composition using MET data. He found no relationship between changes in observation-based practice ratings and changes in classroom composition, consistent with results showing that teacher value added is not affected by classroom composition and with an interpretation that teachers are able to teach at the same level across classrooms of different compositions. However, Steinberg and Garrett (2016), also using MET study data but a different form of analysis, found that English language arts (but not math) teachers working with higher-achieving students "tend to receive higher performance ratings, above and beyond that which might be attributable to aspects of teacher quality that are fixed over time" (p. 20). This is consistent with an interpretation that it is more difficult to teach in ways promoted by the rubrics when assigned a classroom with lower average prior student achievement (although it is puzzling that the effect is more apparent for English language arts teaching compared to mathematics, given that the rubrics used were not subject specific). No studies were located that directly addressed whether particular forms of teaching behaviors are more difficult to carry out with low-achieving students.

Rater Bias

Rater bias, the tendency of raters to be influenced by nonperformance factors when rating, has also long been recognized as a problem (Wherry & Bartlett, 1982). Raters may be inclined to perceive or believe that teachers of disadvantaged students are not teaching as described at higher rubric levels, even if they are. Considerable research on race/gender bias in performance appraisal, although focusing on the characteristics of the ratee, suggests that rater stereotypes can bias ratings (Roberson, Galvin, & Charles, 2007). Stereotype-based bias is suggested by research showing that student characteristics can bias teachers' perceptions of students (e.g., Mason, Gunersel, & Ney, 2014; McGrady & Reynolds, 2013; Riley, 2014). It is possible that observers of teaching could also be biased as to what they observe students doing. They could interpret student behavior as more problematic than the teacher does and consequently rate the teacher as having lower classroom management performance. Or they could hold the stereotype that poor, non-White, or lower-prior-achievement students have difficulty with higher-level thinking skills and thus fail to observe teachers' success in using thought-provoking questions or promoting higher-order thinking skills, as mentioned in evaluation rubrics, like the Framework for Teaching (Danielson, 2007). Coupled with the commonly found tendency for confirmation bias in data collection and judgment (Nickerson, 1998), this could lead to situations in which a rater who believed the students were not prepared to use higher-order thinking skills, or are likely to be less disciplined, overlooks instances of teaching that disconfirm these assumptions and focuses on instances that are consistent with them. It is also possible that school administrators may assign less able students to teachers they perceive are less able. Their perception that the teacher is less able may provide an initial hypothesis that they are subsequently tempted to confirm by ignoring instances of high-performance behavior.

Rubric Deficiency

There are two related forms of rubric deficiency that could contribute to lower ratings for teachers of disadvantaged students. First, teaching behaviors described at lower levels of the rubric could be more effective with disadvantaged students than practices described at the highest levels. If these students do better when teachers use these practices, then teachers who are teaching appropriately would be penalized with lower ratings. Second, and perhaps more likely, rubrics could fail to describe behaviors that are as or more effective with these students than the behaviors described at the highest rubric level. If teachers of disadvantaged students engaged in such behaviors in preference to those described at higher rubric levels (presumably because they have found them effective), raters might assume that the observed behavior best fits a lower rubric level because the behaviors described at the highest level were not observed. In these cases, teachers of disadvantaged students could tend to receive lower ratings than equally effective teachers of other students.

For example, many of the highest levels of the Framework for Teaching rubric (Danielson, 2007) describe teaching in which students take responsibility for aspects of their own learning. For example, Component 3b, "Using Questioning and Discussion Techniques," differentiates between the proficient and distinguished levels of its three elements by referencing student involvement (e.g., "Students formulate many questions"; "Students assume considerable responsibility for the success of the discussion, initiating topics and making unsolicited contributions"; and "Students themselves ensure that all voices are heard in the discussion"). Although I could find no studies that showed that this sort of teaching was not effective for disadvantaged students, it is possible that it could be counterproductive if students lacked enough prior knowledge and interest in the content to keep their contributions relevant to the learning task.[3] Students performing below grade level or with some cognitive disabilities may also require more direct and explicit instruction. For example, it has been claimed that more structured or even scripted lessons would be more effective for

lower-achieving students (e.g., Slavin & Madden, 1987; Slocum, 2004), although the evidence on that claim is mixed (e.g., Borman et al., 2005; Gersten, 1985; Ross et al., 2004; What Works Clearinghouse, 2007). It has also been argued that standard rubrics may not recognize specific special education pedagogy (e.g., Council for Exceptional Children, 2012; Johnson & Semmelroth, 2014; Woolf, 2015) and that commonly used rubrics, like the Framework for Teaching, may not be valid for evaluation of special education teachers (Jones & Brownell, 2014).

Another way rubrics could be deficient is that they may not provide enough guidance for raters to understand how to apply higher levels in different situations (such as with different kinds of students). For example, although it may be possible to provide students with cognitive or behavioral disabilities with opportunities to take some responsibility for their own learning or discipline, this might look so different in a special education classroom that the rater, considering just the explicit rubric language, could have trouble recognizing it. Although rubrics cannot describe all variation without becoming unwieldy, some rubrics may be better than others in describing behavior in ways that allow raters to see how the concepts apply to different situations.

Like rater bias, rubric deficiency does not directly influence teacher behavior but influences how observed behavior is translated into ratings. However, unlike the effects of rater bias, the ratings made using these rubrics can accurately represent observed behavior. The teachers affected are not behaving in ways that justify a higher rating based on the rubric as written, although the rubric does not capture all effective behaviors. Rubric deficiency, if it exists, seems unfair, however, because these teachers would be penalized for using effective behaviors or, at worst, have to choose between behaviors that improve student achievement and getting a higher rating.

How Would Ratings Be Adjusted?

In order to remove the effects of observed classroom composition from the ratings, the obvious course is to control for student characteristics in much the same way as is done in value-added modeling. In a typical value-added model, students' test scores are modeled as a function of prior test scores (in the same and sometimes other subjects as well) to control for prior knowledge and ability; a set of student characteristics, such as free and reduced-price lunch eligibility, race, special education status, whether the student was an English learner, and sex; and in some cases, other available information, such as whether a student is gifted or homeless. Some models also include controls for classroom composition, in the form of average prior-year scores or percentages of students with characteristics expected to be correlated with test scores. Teacher effects are represented by coefficients for indicator variables included for each teacher (in fixed-effects models) or by the average of residuals for each teachers' students (in random-effects models).

When used to estimate teachers' effects on student achievement, controls for student characteristics are intended to remove alternative causes of student test scores, allowing the remaining to be attributed to the teacher. Whether value-added models succeed in producing unbiased estimates of teacher effects on student achievement is controversial. The major issue is whether nonrandom sorting of students to teachers leads to imbalances in unobserved influences of student achievement that advantage some teachers and disadvantage others (AERA, 2015; Berliner, 2014; Rothstein, 2010). The types of student characteristics commonly included in value-added model specifications likely do not fully account for influences like student motivation and interest, parental engagement, or summer learning loss. The assumption needed to attribute impacts to teachers is that the differences between teachers' classrooms in these unobservables that are not absorbed by the controls are random and would average out over large samples of students and teachers. Although the controversy about bias in value-added models is ongoing, much of the research on its size suggests it is fairly small (e.g., Bacher-Hicks, Chin, Kane, & Staiger, 2015; Chetty, Friedman, & Rockoff, 2014; Koedel, Mihaly, & Rockoff, 2015). Although some level of sorting bias is likely to exist, estimates from models including student characteristics are less likely to show relationships with student characteristics than from models that do not, suggesting that when one turns to rating teacher practice, controlling for these is more accurate and fair than not doing so.

Applying this strategy to practice ratings would involve calculating a residual from a predicted evaluation rating based on regressing raw ratings on various controls for classroom composition, such as average prior student achievement, proportion of students eligible for free or reduced-price lunch, and/or proportion of students who are English learners. Teachers would have their scores adjusted in proportion to the relative disadvantage of their classrooms and the size of the coefficients for the classroom composition measuring disadvantage included in the model. For example, a rating of 2.80 could be adjusted up to 3.0 for a teacher with a classroom with more disadvantaged students than average, and a rating of 3.0 could be adjusted down to 2.80 for a teacher whose students are less disadvantaged than average, where a rating of 3.0 was considered proficient practice as defined by the rubrics. Although this adjustment could be done at the dimension (e.g., Component 3a of the Framework for Teaching, "Communicating With Students") or domain level (e.g., "Instruction" in the Framework for Teaching), the final summative rating would more likely be adjusted, because this is what is typically used for tenure, retention, or performance pay decisions.

Although this adjustment could remove the correlation between ratings and classroom composition, there are some reasons to be cautious about the analogy between using value-added models to produce less-biased estimates of teacher impacts on student achievement and controlling for classroom composition to better measure teacher practice.

First, adjustment would change the meaning of ratings, obscuring the link between the ratings and the underlying rubrics. Current rubrics are standards based, much like the tests used to assess student performance.[4] They are designed to represent a developmental progression of performance, up to and past a level (e.g., "proficient") that all teachers are expected to meet, and are intended to measure teacher performance against an ideal standard and hold teachers accountable for teaching to it. Adjustment would remove much of the instructional guidance function of ratings based on rubrics. The teacher whose unadjusted rating was 3.0 could interpret that rating as indicating that practice was "proficient" on average. If adjusted to 2.8 due to classroom composition, the adjusted rating indicates below-proficient performance as compared not to the rubric but rather to the performance of teachers with the average classroom composition. Further, because performance is now defined relative to that of other teachers, a teacher cannot be sure that improving performance will improve her or his rating. If other teachers improve as well, this teacher could find that her or his rating has stayed the same or even declined. Last, there is some evidence that changing the focus to interpersonal comparison could be counterproductive to using ratings to motivate improvement (Anseel, Van Yperen, Janssen, & Duyck, 2011; DeNisi & Kluger, 2000; Luffarelli, Gonçalves, & Stamatogiannakis, 2016). At the least, both adjusted and unadjusted score would have to be provided to teachers if they are expected to use the results to improve their practice.

Second, teachers are likely to have greater control over their own practice than they do over student achievement. The causal link between teachers' abilities, skills, beliefs, and level of effort and classroom behaviors is more direct than their link to student achievement. In value-added analyses, controls for student characteristics are expected to remove important influences on test scores teachers cannot control, whereas in most cases teachers are expected to adapt their practice to their students. Adjustment reduces the incentive to adapt and our ability to assess how well teachers succeed. Although fairness is important, the trade-off between potentially overcontrolling for classroom composition and not controlling enough could be different when the intent is to influence the mechanism by which teachers influence achievement. Although overcontrolling is a possibility in value-added estimation (Ballou et al., 2004; McCaffrey, 2012), it is more likely in adjusting ratings.

Third, teachers' behavior is much more observable than teachers' effects on student achievement. When estimating the latter, there are few practical alternatives to using observable student characteristics as proxies to control for unobservables, like student motivation, engagement, parental influences, and other actual causes of student achievement outside teachers' control. But because behavior can, with care, be observed, there is no need to control for classroom composition to get an accurate measure of the teacher's practice. If it is harder for teachers to behave as expected the more disadvantaged their students, it might be fairer to adjust ratings, but that would depend on the intended uses, as discussed in the next section.

To Adjust or Not to Adjust?

When considering the appropriateness of adjustment when teachers of disadvantaged students are found to have lower ratings, the cause of the difference and the use to be made of the rating need to be considered together. Different uses imply different inferences about what the ratings mean, and different conclusions about whether the validity of the inference would be enhanced or reduced by adjustment. Table 1 summarizes recommendations for adjustment for each combination of the five potential causes discussed above and five typical uses of performance ratings.

As shown in the second column of Table 1, when differences in ratings between teachers of disadvantaged and other students are due to differences in teacher behavior that are rooted in differences in skills or beliefs, adjustment is not recommended regardless of purpose. In these cases, there are true differences in behavior that adjustment would mask. Adjustment would communicate that teachers of disadvantaged students are held to a lower standards of skill, distort feedback by implying that poorly performing teachers have less improvement to make, and miss some teachers on the margin of benefiting from professional development. Teachers whose skills are not sufficient to perform to the standard are not likely to be the ones that a district would want to receive job protections, given the difficulty of later removing teachers for performance. Most districts would not want to retain these teachers in preference to others or spend scarce compensation resources on them.

If ratings differentials are caused by greater difficulty in teaching disadvantaged students, and ratings are to be used for setting a standard of practice, providing feedback, identifying teachers for professional development, or ensuring disadvantaged students receive the same quality of instruction as others, adjustment for classroom composition would mask true differences in practice that users would be trying to reduce. Adjustment for classroom composition would set lower standards for teachers of disadvantaged students, distort feedback, and miss teachers who could benefit from professional development. In contrast, if ratings are to be used for employment or pay, adjustments for classroom composition could improve the validity of inferences and

TABLE 1
Recommendations for Adjustment of Ratings by Source of Rating Difference and Use

Use	Potential cause of differences in ratings				
	Teacher skills, abilities, attitudes	Student disadvantage	Teaching conditions	Rater bias	Rubric deficiency
Communicating a standard of practice and providing feedback to teachers	No adjustment	No adjustment	No adjustment	No adjustment; retrain raters	No adjustment
Identifying teachers for professional development	No adjustment	No adjustment	No adjustment	Adjust for rater severity	No adjustment
Ensuring equitable access	No adjustment	No adjustment	No adjustment	Adjust for rater severity	Change rubric
Making probation, tenure, or retention decisions	No adjustment	Adjust for composition	No; equalize conditions	Adjust for rater severity	Change rubric
Performance-based compensation	No adjustment	Adjust for composition	No; equalize conditions	Adjust for rater severity	Change rubric

reduce disincentives for teaching disadvantaged students. The inferences typically behind these uses go beyond that teachers exhibited specific behaviors. Teachers are given tenure or retained because it is believed that they will be at least acceptable performers in the future. Although performance pay is presented as a reward for past performance, its function is to motivate future performance or retain teachers likely to perform well in the future. The underlying inference is that the rating represents likely future performance, and its validity depends in part on whether classroom composition and teaching conditions would be the same in the future. If teachers currently teaching disadvantaged students could be assigned different students in the future, an adjustment could improve the prediction of future performance by estimating what the rating would be if the teacher were assigned to an average classroom. If differences in ratings are large enough to materially affect chances of tenure, nonretention, or pay increases, teachers may be less likely to teach under conditions that make achieving the needed rating more difficult. A properly designed adjustment could equalize the chances of receiving these benefits and reduce the disincentive.

If ratings differentials are caused by differences in teaching conditions, unless classroom composition is strongly correlated with teaching conditions, adjusting for it will not address differentials due to teaching conditions and could mask low-quality teaching by teachers with better-than-average teaching conditions but more-than-average disadvantaged students. Further, unless the correlation is high, adjusting for classroom composition here would be unfair to teachers who have poor teaching conditions but less disadvantaged students. Adjusting for teaching conditions themselves is likely to be complex. For example, does the difference between a class of 20 in math and a class of 30 in art, band, or physical education disadvantage the latter teachers? It would probably be simpler in the long run to try to equalize class sizes or resource levels across comparable teachers than to find a fair method of measuring and adjusting for teaching conditions. Even if more difficult teaching conditions are highly correlated with student disadvantage, fairness to the students requires equalizing conditions.

Rater bias differs from the other causes in that teachers of disadvantaged students could be teaching as intended, but this is being masked by raters' failure to recognize it. Rater bias is the only cause that fits the technical definition of measurement bias, as a difference between an obtained measurement and its true value. Adjustment seems the most unambiguously appropriate here, because if the bias can be removed by adjustment, ratings would better represent actual behavior, and inferences about which teachers should be trained, retained, or rewarded would be more valid. However, adjusting based on classroom composition would not be justified unless the vast majority of raters exhibited a similar degree of bias. If bias associated with classroom composition varied among raters, the more appropriate method of adjustment would be based on individual raters' degree of bias.[5] Further, if the primary use is to communicate a standard of practice and provide feedback to teachers, any form of adjustment would be less useful than removing the rater bias by retraining raters or selecting less biased ones, because adjustment makes it harder for teachers to compare their ratings to the rubric.

Where differences are caused by rubric deficiency, adjustment would be counterproductive for using the rubrics as standard of practice, providing feedback about how well that standard has been met, and monitoring whether disadvantaged students experience teaching that meets the standard. Adjustment would also reduce usefulness for identifying which teachers need professional development to teach according to the rubrics. Although adjustment for purposes such as making probation, tenure, retention, or compensation decisions could improve fairness, it risks

covering up situations where teachers of disadvantaged students are using neither the behaviors promoted by the rubrics nor more or equally effective behaviors. The correlation between the use of alternative, but equally or more effective, behaviors and classroom composition would have to be quite high to avoid overadjustment. Modifying the rubric would be more appropriate. One approach could be to modify them for specific contexts, as has been done for special education (Holdheide, 2013). Modifying them for specific subjects might also improve their usability for recognizing important but subtle differences in instruction (Hill & Grossman, 2013).

Researching the Causes of Rating Differentials

To better inform decisions about adjusting ratings when differences associated with student disadvantage appear, additional research is needed aimed at assessing which of the potential causes operate and their relative importance. One way to think about such research is by analogy with generalizabity theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Four of the five potential causes discussed above can be seen as potential facets of measurement error: raters, classroom composition, teaching conditions, and rubrics. Teacher skills, abilities, and beliefs are indirectly measured by the ratings we would like to generalize about across the other facets. If the appropriate randomization and other requirements could be achieved, the variance disaggregation for the various main effects and interactions could provide estimates of the relative importance of these effects.

Whereas a pure generalizability study design is likely impractical, a more limited study could be done. It would be essential to randomly assign teachers to classrooms with varying compositions in order to separate the effects of teacher skills, abilities, and beliefs from the effects of classroom composition. Ideally, teaching conditions would also be randomly assigned, but this is impractical and likely unethical, so one would have to settle for existing variation in conditions found in the classrooms assigned to teachers and the possibility of not being able to separate working conditions from classroom composition.

In a typical evaluation setting, teachers are rated by a school administrator who evaluates teachers with similar classroom compositions. To disentangle raters from classroom composition, it would be desirable to randomly assign a second rater. This would allow an assessment of the effects of classroom composition on raters as well as comparing ratings by someone without an ongoing relationship with the teacher to ratings made by administrators.

Using multiple rubrics and randomizing rubrics to raters could enable assessing the potential impact of rubric deficiency. The variation in the rubrics should reflect a hypothesis about the teaching behaviors that might be effective with disadvantaged students that are not well captured by frequently used rubrics. Otherwise, the differences in ratings across rubrics will likely be due to differences in content or wording unrelated to the type of rubric deficiency of interest.[6]

If these requirements are met, ratings could be modeled as a function of classroom composition, teaching conditions, rater, rubric, and interactions between raters and composition and rubric and composition. The main effect of classroom composition would represent the effect of the difficulty of teaching specific types of students (but possibly confounded with teacher beliefs, like lower expectations for poor or lower performing students). The main effect of teaching conditions would represent the effect of the difficulty of teaching under the specific types of conditions. The interactions between rater and composition would represent the bias of individual raters related to classroom composition; and the average of these interactions, the average rater bias related to composition. The interaction between rubric and composition would represent rubric deficiency related to classroom composition. Other interactions that could be of interest if they could be estimated with precision could include teachers with composition, representing differential ability to teach according to the rubric in classes with different compositions, analogous to differential teacher effectiveness in value-added analyses (Lockwood & McCaffery, 2009; Meyer & Dockumaci, 2015).

Because randomization is likely to be difficult to achieve, it would also be valuable to collect multiple years of data.[7] This would allow the inclusion of teacher fixed effects, as in Steinberg and Garrett's (2016) study, providing an estimate of persistent skills and abilities of teachers and controlling of those effects when estimating the effect of classroom composition. Although having multiple years of data allows less reliance on randomization, by itself it may not be sufficient to separate the effects of teacher skills and abilities from student composition, unless the proportions of poor or low-achieving students assigned to teachers change over time. This would be less likely in districts where teachers have low interschool mobility and where schools' classroom composition is homogeneous.

Observations of teaching and interviews with teachers and raters also have an important role to play in studying the causes of ratings differentials. Quantitative analyses of ratings cannot differentiate between lower ratings due to lower teacher expectations for disadvantaged students and differences in the difficulty of teaching them. One way of trying would be to observe, interview, and analyze related artifacts (e.g., student work) from teachers of high-poverty/low-prior-achievement classrooms who get better-than-expected ratings given their classroom composition and compare what they do and how they think about teaching their students to teachers who do just as expected. Teachers could be identified for study by plotting ratings by classroom proportion of

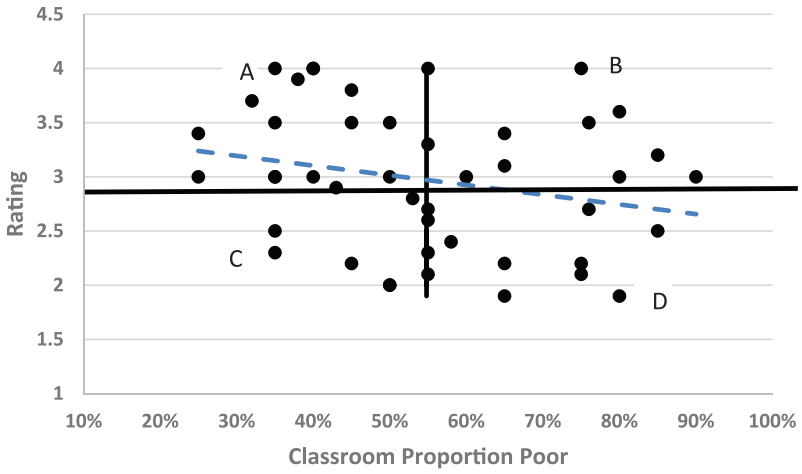

FIGURE 2. *Plot of hypothetical practice ratings by classroom poverty.*

poor students, as in Figure 2. Referencing the figure, if teacher expectations were an important reason for rating differences, one would expect to hear more responses and see more behavior indicative of low expectations when moving downward on the right side of the average proportion of classroom poverty line from B to D. One would expect to hear and see more related to skill issues or indicative of difficulty of teaching moving down from A to C to the left of the average classroom poverty line.

Observing and interviewing a sample of these teachers would also allow confirming that they were appropriately rated and whether those rated above average were using teaching techniques that are hypothesized to be related to teacher expectations. If so, their perceptions of their own skills, their students' capacities, and their expectations of what teaching techniques could be used can be contrasted to those of teachers with similar proportions of disadvantaged students that do not use the techniques. If the reason for not using techniques associated with higher ratings is lower expectations, one would expect to hear teachers not using them talk about expectations rather than skills as the rationale and to hear teachers who do use them express confidence that the techniques can be used with their students. Of course, this approach would require the employment of skilled observers/interviewers who both understand teaching and can gain the trust of the teachers being studied.

Observations and interviews would also be useful in studying rater bias. As argued above, it is important to determine whether rater bias is a general tendency found in all or most raters or whether the bias varies substantially among the raters of teachers of disadvantaged students. The individual rater–composition interactions in the analysis proposed above would provide evidence as to how similar bias is for different raters. If the degree of bias appears to differ across raters, studying the decision-making processes of raters with different degrees of estimated bias who evaluated teachers with similar classroom compositions could help determine what raters are missing and how rater training can be improved. Closer study of rater decision making could also determine whether raters might be *more* lenient in schools serving disadvantaged students. These raters may tend to make excuses for poor teaching, or may be reluctant to discourage teachers from staying by giving low ratings, fearing that these teachers will be hard to replace. It is also possible that if high-ability teachers are less common in these schools, the raters (administrators) may rarely see really good teaching and assume that the teaching they see is at least proficient. Such leniency could lead to an underestimate of the difficulty teachers of disadvantaged students could have in teaching in ways that merit high ratings.

Closer study of teacher practice could also identify rubric deficiency. Interviews and observations of teachers of disadvantaged students could uncover whether they are using practices that are just as effective as or more so than those described in the rubrics. Teachers who appear to be effective with disadvantaged students but receive lower-than-expected practice ratings can be identified using value-added methods. These would be teachers of disadvantaged students with higher-than-expected value-added scores but lower practice ratings than teachers of other students with similar value-added scores. These teachers would be found in the shaded area shown at the top right of Figure 3, which plots hypothetical value-added estimates of effectiveness against the percentage of disadvantaged students. If practices are being used that are effective with disadvantaged students but are not highly rated, then they should be discernable from studying these teachers. Studying the observation and decision-making practices of their raters would help determine if practices equivalent to those described at higher rubric levels were being overlooked by raters.

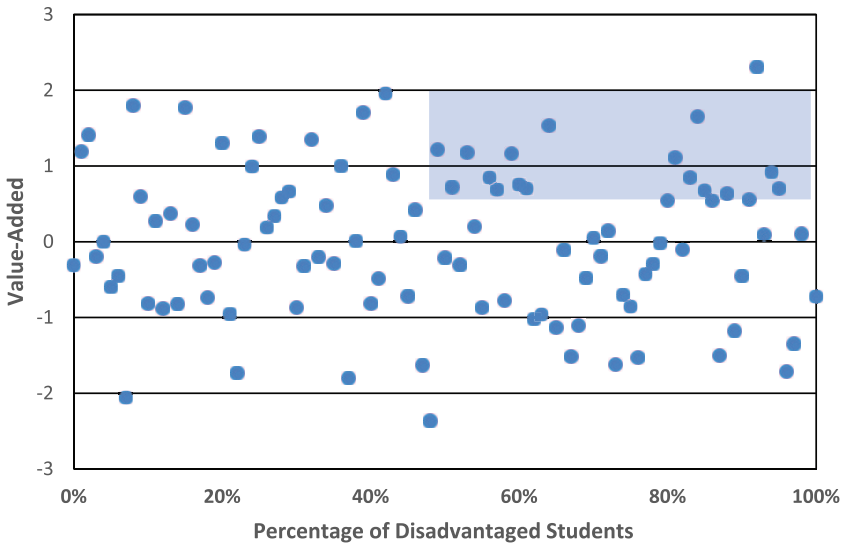

FIGURE 3. *Hypothetical set of teachers with high value added and high percentage of disadvantaged students likely to be using effective practices unrecognized by practice rubrics.*

Discussion

This article has argued that there are several potential causes of differences in performance evaluation ratings between teachers of disadvantaged students and teachers of their more advantaged peers, including differences in teachers' skills, abilities, and beliefs, in the difficulty of teaching students of different degrees of disadvantage, teaching conditions, rater bias, and rubric deficiency. It further argued that whether to adjust teacher evaluation ratings for classroom composition, in an attempt to be fair to teachers of disadvantaged students, should be considered in light of which of the potential causes are at work and the uses to which the ratings will be put. Some combinations of causes and uses do not justify simply adjusting ratings for classroom composition. In particular, rater bias and rubric deficiency should be addressed by training or changing raters or modifying rubrics. These are particularly pernicious—the former because it disguises what is really happening in the classroom and the latter because it distorts the incentives for teaching in ways that benefit students that evaluation systems are intended to provide.

Although further research could help determine which causes are most important, it is likely that multiple causes will be found, and many policymakers do not have time to wait for the needed research to accumulate. In addition, most ratings are intended to have multiple uses. Adjusting for classroom composition is simpler than pursuing other causes, and although imperfect, the analogy with value added is likely to make it attractive. The decision to adjust for classroom composition may come down to a trade-off between the potential negative effects of covering up true differences in teaching and making ratings harder to use to improve it, on one hand, and improving perceptions of fairness and lowering disincentives to teach disadvantaged students, on the other.

If the major use is to track teacher performance and provide feedback to improve it, then adjustment for classroom composition has the potential to obscure needs for improvement and hinder efforts to ensure poor or lower-performing students have access to the kind of instruction the evaluation system is intended to promote. The experience with value added is instructive here. Although most methodologists accept the need for adjusting for poverty and race in value added, these adjustments have apparently put the problem of lower adjusted achievement being associated with poverty or race out of mind. Few if any jurisdictions appear to track whether teachers or schools are reducing the size of the negative coefficients on poverty or race.

If the most important use is to make between-teacher comparisons or rankings for high-stakes decisions, the case for adjustment is much stronger. If teachers are not given similar students to teach, and the difficulty of teaching different types of students differs substantially, comparing teacher performance would be unfair, and the inference that a teacher with higher ratings but higher-prior-achieving or less poor students is a better teacher than one with lower ratings but poorer or lower-achieving students could be invalid. These uses of unadjusted ratings might also be counterproductive. Teachers appear to be quite sensitive to perceived unfairness in evaluation, and, especially if financial consequences stem from evaluation ratings (e.g., Goldhaber, Bignell, Farley, Walch, & Cowan, 2014; Heneman & Milanowski, 2003), disincentives to teach disadvantaged students would only be strengthened. Adjusting ratings could potentially improve perceptions of fairness, at the expense of making ratings less transparent for other uses.

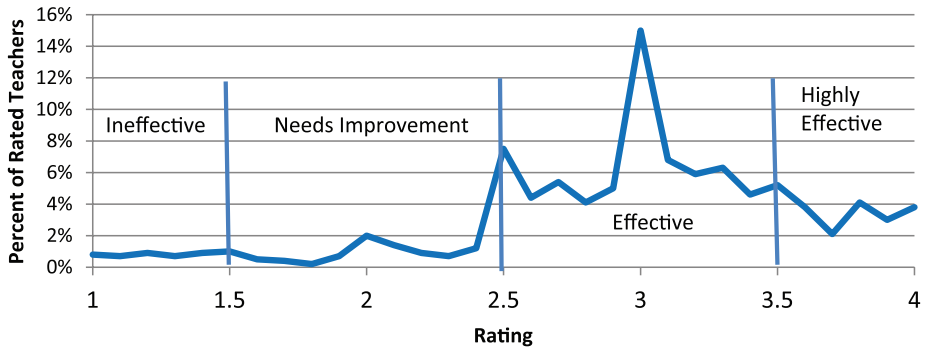

FIGURE 4. *Hypothetical distribution of ratings and cutoff points for performance categories.*

Although adjustment could improve fairness and accuracy in ranking, many states and districts do not actually rank teachers but rather "bucket" them into broad categories of effectiveness or performance, such as ineffective, needs improvement, effective, and highly effective. Figure 4 shows a hypothetical distribution of teacher ratings and common cut points used to divide the rating range into effectiveness categories. As the figure illustrates, most teachers' ratings are not near the cut points, and so a small difference between adjusted and unadjusted ratings is unlikely to make a difference in the outcomes of the evaluation except to teachers near the cut points. Rather than adjusting ratings, it could be simpler and more transparent to review the practice of teachers of disadvantaged students whose ratings are close to the cutoff points before making consequential decisions based on those ratings.

If evaluation ratings are to be used as a sort of skill test for teachers, to represent their long-term and generalizable ability to deliver quality instruction, differences in the difficulty of teaching students introduce unfairness, invalidate comparisons, and compromise generalizability. Yet simple adjustment for classroom composition may not go far enough. As Reckase and Martineau (2015) pointed out in the context of value added, if the goal is to estimate locations of teachers on a latent construct of teaching capability, teachers should be "tested" with students of varying difficulty that function like the differing items of a test. If student difficulty cannot be equated by equalizing classroom composition, adjustment would be justified, but Reckase and Martineau's arguments imply that adjustment might have to go further, including adjusting for rater leniency, class size, and potentially myriad other influences. But this would further reduce transparency.

The more one thinks about evaluation as a test, the more worried one becomes about explicitly controlling for or standardizing factors that could affect performance. All or some of these factors may be associated with differences in ratings. Once adjustment begins, where should it stop?[8] More extensive adjustment may also be confounded with subjective adjustments raters already make to account for situational constraints on performance (such as class size or even the perceived difficulty of teaching different types of students), adjustments that likely vary by rater (Dobbins, Cardy, Facteau, & Miller, 1993; Kane, 1997). Daunted by the difficulty of ensuring the validity of performance measures based on judgments of complex performance, some organizational psychologists have moved from emphasizing measurement quality to making evaluation more useful in improving performance (e.g., DeNisi & Pritchard, 2006; Murphy, 2008). There is growing recognition that pursuing high levels of validity may be unrealistic and of lower priority than helping organizations figure out how to use the evaluation process to improve employee performance. Teacher evaluation may have to move in a similar direction.

Concerns about measurement validity appear to have developed with proposals for high-stakes uses of evaluation results. When evaluation was primarily used for building a case for terminating the worst performers, measurement concerns were muted. Such limited consequential use, coupled with using ratings to provide feedback and plan professional development, could be attractive to many policymakers (e.g., Kimball, Rainey, & Mueller, 2016) because it avoids the need to prove the validity of ratings and reduces potential conflict. Although this approach could be viable if policymakers and practitioners were serious about linking the evaluation process to professional development, it does forgo using evaluation results to change the composition of the teaching workforce by differentially retaining higher-rated teachers. It has long been known that using measures of even moderate validity for selection and deselection can improve overall workforce quality (Taylor & Russell, 1939). What is not yet known is whether a concerted use of evaluation results to improve performance in a low-stakes environment can work at scale.

If the proposed distinctions between causes of ratings differences discussed here are credible, until more is known about which are dominant, it is premature to label evaluation systems producing ratings that differ with classroom composition as biased or unfair. If teachers of disadvantaged students are not using techniques that are described at higher

levels of evaluation rubrics while teachers of other students are, ratings differences are not due to bias. Nor is it unfair that teachers of these students are rated lower if they are less skilled or hold low expectations for their students. The observed differences do indicate the need for studies to examine their causes. Until those studies are done, is it justified to use ratings for consequential decisions? There is no easy answer. Like most decision processes that affect both individuals and the broader organization, there is a trade-off between the cost to individuals from using an imperfect process and the cost to the organization and the people it serves of not using it. We need to know more about the size, prevalence, and causes of differences in ratings, and the benefits of their uses, in order to make this trade-off.

Notes

1. Throughout the article, *disadvantage* is used as a shorter way to refer to student characteristics associated with lower ratings or lower student achievement, including prior achievement and poverty but also possibly including race, special education needs, and being an English learner.

2. Walsh and Lipscomb (2013) found that a substantial proportion of variance in practice domain ratings was between schools in Pennsylvania. This is consistent with the possibility that some of the differences in ratings might be due to school-level conditions that are correlated with student poverty or average schoolwide achievement. For rating differences due to these factors to exist within districts, schools would need to vary substantially within districts on class size or resource levels.

3. Like other rubric designers (e.g., Marzano, 2007; Pianta, LaParo, & Hamre, 2008), Danielson (1996) cites a variety of research supporting the concepts underlying her rubric. In addition, the efficacy of the constructivist techniques described in the higher levels of the Framework for Teaching draws plausibility from research on how people learn (e.g., Bransford, Brown, & Cocking, 1999, 2000).

4. Although the tests are standards based, estimating teacher impacts on student achievement requires a relative standard, because no one yet knows what the appropriate size of a teacher's effect should be. In contrast, teaching practice rubrics purport to provide an absolute standard of teaching.

5. Such adjustments can be made based on analyses of ratings done using variants of item response theory (e.g., Kelcey, McGinn, & Hill, 2014; Wolfe, 2004). They would likely require that at least some teachers be rated by multiple observers.

6. Although randomizing the use of multiple rubrics is likely to be challenging, it would be possible to explore rubric effects in a more limited way by video recording a sample of observations representing classrooms with different levels of student disadvantage, then evaluating them on different rubrics by sets of expert observers, each set trained on one of the alternative rubrics. This was done in the Measuring Effective Teaching (MET) study (Kane & Staiger, 2012). Although observer effects cannot be separated from rubric effects, over a big enough sample of teachers and observers, one would get some information as to whether using one rubric was more favorable to teachers of disadvantaged students than the other. Another approach would be to identify behaviors effective with disadvantaged students from studying teachers shown to be effective with them and seeing if observers rating videos using the rubric under study (for example, a general rubric, like the Framework for Teaching) missed these behaviors.

7. As Steinberg and Garrett (2016) observed, randomization was incomplete in the MET study, despite the incentives districts received to randomize from the Gates Foundation.

8. For example, grade level could also affect ratings. Lazerev and Newman (2015) found that the relationship between ratings and classroom characteristics was larger in the highest grade they studied (eighth) and smaller in the lowest (fourth). It is possible that it becomes more difficult to teach lower-achieving students in ways promoted by the rubrics as they get older, and it is also possible that teacher expectancy has more scope to operate. The performance of peers may also affect teacher performance (Jackson & Bruegmann, 2009), as might the quality of school leadership (Heck, 2012).

References

Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The schools teachers leave: Teacher mobility in Chicago public schools*. Chicago, IL: Consortium on Chicago School Research.

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, *44*(8), 448–452.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Anseel, F., Van Yperen, N. W., Janssen, O., & Duyck, W. (2011) Feedback type as a moderator of the relationship between achievement goals and feedback reactions. *Journal of Occupational and Organizational Psychology*, *84*(4), 703–722.

Bacharach, S. B., & Bamberger, P. (1995). Beyond situational constraints: Job resources inadequacy and individual performance at work. *Human Resource Management Review*, *5*(2), 79–102.

Bacher-Hicks, A., Chin, M., Kane, T. J., & Staiger, D. O. (2015, March). *Validating components of teacher effectiveness: A random assignment study of value-added, observation, and survey scores*. Paper presented at the spring 2015 conference of Society for Research on Educational Effectiveness, Washington, DC.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record*, *116*(1), 31.

Blatchford, P., & Mortimore, P. (1994). The issue of class size for young children in schools: What can we learn from research? *Oxford Review of Education*, *20*(4), 411–428.

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and education equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *Elementary School Journal*, *106*(1), 3–20.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005). The national randomized field trial of success for all: Second-year outcomes. *American Educational Research Journal*, *42*(4), 673–696.

Bourke, S. (1986). How smaller is better: Some relationships between class size, teaching practices, and student achievement. *American Educational Research Journal*, *23*(4), 558–571.

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2010). *The role of teacher quality in retention and hiring: Using applications-to-transfer to uncover preferences of teachers and schools* (NBER Working Paper No. 15966). Cambridge, MA.National Bureau of Economic Research.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How students learn: History, mathematics, and science in the classroom*. Washington, DC: National Academy Press.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh public schools* (REL 2014-024). Calverton, MD: Regional Educational Laboratory Mid-Atlantic.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Chingos, M. M., & Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review*, *30*(3), 449–465.

Clotfelter, C., Ladd, H. F., & Vigdor, J. L. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*(4), 377–392.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119–142.

Cohen, J., & Goldhaber, D. (2016). Observations on evaluating teacher performance. In J.A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 8–21). New York, NY: Teachers College Press.

Council for Exceptional Children. (2012). *The Council for Exceptional Children's position on special education teacher evaluation*. Arlington, VA: Author.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., . . . Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *Elementary School Journal*, *112*(1), 16–37.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

DeAngelis, K., Presley, J., & White, B. (2005). *The distribution of teacher quality in Illinois* (Policy Research Report IERC 2005-1). Edwardsville, IL: Illinois Education Research Council.

DeNisi, A. S., & Kluger, A. N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? *Academy of Management Executive*, *14*, 129–139.

DeNisi, A. S., & Pritchard, R. D. (2006). Performance appraisal, performance management, and improving individual performance: A motivational framework. *Management and Organization Review*, *2*(2), 253–277.

Dobbins, G. H., Cardy, R. L., Facteau, J. D., & Miller, J. S. (1993). Implications of situational constraints on performance evaluation and performance management. *Human Resource Management Review*, *3*(2), 105–128.

Figlio, D. N. (2007). Boys named Sue: Disruptive children and their peers. *Education Finance and Policy*, *2*(4), 376–394.

Fox, L. (2016). Playing to teachers' strengths: Using multiple measures of teacher effectiveness to improve teacher assignments. *Education Finance and Policy*, *11*(1), 70–96.

Gersheson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me: The effect of student-teacher demographic match on teacher expectations. *Economics of Education Review*, *52*, 209–224.

Gersten, R. (1985). Direct instruction with special education students: A review of evaluation research. *Journal of Special Education*, *19*(1), 41–58.

Goldhaber, D., Bignell, W., Farley, A., Walch, J., & Cowan, J. (2014). *Who chooses incentivized pay structures? Exploring the link between performance and preferences for compensation reform in the teacher labor market* (CEDR Working Paper 2014-8). Seattle, WA: Center for Education Data and Research.

Graue, E., Hatch, K., Rao, K., & Oen, D. (2007). The wisdom of class-size reduction. *American Educational Research Journal*, *44*(3), 670–700.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2001). Why public schools lose teachers. *Journal of Human Resources*, *29*(2), 326–354.

Harris, D. N., & Sass, T. R. (. (2009). What Makes for a Good Teacher and Who Can Tell? Working Paper 30. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Heck, R. H. (2012). Instructional practice, teacher effectiveness, and growth in student learning in math: Implications for school leadership. In B.G. Barnett, A. R. Shoho, & C. A. Tooms (Eds.), *The changing nature of instructional leadership in the 21st century* (pp. 33–62). Charlotte, NC: Information Age.

Heneman, H. G., III, & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, *17*(2), 173–195.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371–384.

Holdheide, L. (2013). *Inclusive design: Building educator evaluation systems that support students with disabilities. Special issues brief* (Rev. ed.). Washington, DC: Center on Great Teachers and Leaders.

Jackson, K. C., & Bruegmann, E. (2009). *Teaching students and teaching each other: The importance of peer learning for teachers* (NBER Working Paper No. 15202). Cambridge, MA: National Bureau of Economic Research.

Jacob, B. A., & Walsh, E. (2011). What's in a rating? *Economics of Education Review*, *30*(3), 434–448.

Johnson, E., & Semmelroth, C. L. (2014). Special education teacher evaluation: Why it matters, what makes it challenging, and how to address these challenges. *Assessment for Effective Intervention*, *39*(2), 71–82.

Jones, N. D., & Brownell, M. T. (2014). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention*, *39*(2), 112–124.

Kane, J. S. (1997). Assessment of the situational and individual components of job performance. *Human Performance*, *10*(3), 193–226.

Kane, T., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Research report. Seattle, WA: Bill & Melinda Gates Foundation.

Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology*. http://dx.doi.org/10.3389/fpsyg.2014.01469

Kimball, S. M., Rainey, K. M., & Mueller, M. (2016). Learning-centered teacher evaluation in Wisconsin. In T. M. Petty, A. Good, & S. M. Putnam (Eds.), *Handbook of research on professional development for quality teaching and learning* (pp. 618–633). Hershey, PA: Information Science Reference.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.

Kolb, B., & Gibb, R. (2015). Childhood poverty and brain development. *Human Development, 58*(4/5), 215–217.

Kunter, M., & Baumert, J. (2006). Who is the expert? construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251.

Lazarev, V., & Newman, D. (2015, MONTH). *How teacher evaluation is affected by class characteristics: Are observations biased?* Paper presented at the annual meeting of the Association for Education Finance and Policy, Washington, DC.

Lipina, S. J., & Colombo, J. A. (2009). *Poverty and brain development during childhood: An approach from cognitive psychology and neuroscience. Human brain development series* Washington, DC: APA Books.

Lockwood, J. R., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*(4), 439–467.

Luffarelli, J., Gonçalves, J., & Stamatogiannakis, A. (2016). When feedback interventions backfire: Why higher performance feedback may result in lower self-perceived competence and satisfaction with performance. *Human Resource Management*, *55*(4), 591–614.

Malmberg, L., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, *102*(4), 916–932.

Marzano, R. J. (2007). *The art and science of teaching*. Alexandria, VA: ASCD.

Mason, B. A., Gunersel, A. B., & Ney, E. A. (2014). Cultural and ethnic bias in teacher ratings of behavior: A criterion-focused review. *Psychology in the Schools*, *51*(10), 1017–1030.

Max, J., & Glazerman, S. (2014). *Do disadvantaged students get less effective teaching? Key findings from recent institute of education sciences studies. NCEE evaluation brief* (NCEE 2014-4010). Jessup, MD: National Center for Education Evaluation and Regional Assistance.

McCaffrey, D. F. (2012). *Do value-added methods level the playing field for teachers? What we know series: Value-added methods and applications* (Knowledge Brief 2). Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

McGrady, P. B., & Reynolds, J. R. (2013). Racial mismatch in the classroom: Beyond Black-White differences. *Sociology of Education*, *86*(1), 3–17.

Meyer, R., & Dokumaci, E. (2015). Value-added models and the next generation of assessments. In R. W. Lissitz (Ed.), *Value added modeling and growth modeling with particular application to teacher and school effectiveness* (pp. 139–190). Charlotte, NC: Information Age.

Milanowski, A. T., Kimball, S., & White, B. (2014, April). *The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 148–160.

Marchitello, M., & Wilhelm, M. (2014). *The cognitive science behind the Common Core*. Washington, DC: Center for American Progress.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23).

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Peters, L. E., & O'Connor, A. (1980). Situational constraints and work outcomes; the influence of a frequently overlooked construct. *Academy of Management Review*, *5*(3), 391–397.

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System manual K–3*. Baltimore, MD: Brookes.

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*(2), 183–183.

Pong, S., & Pallas, A. (2001). Class size and eighth-grade math achievement in the United States and abroad. *Educational Evaluation and Policy Analysis*, *23*(3), 251–273.

Protik, A., Walsh, E., Resch, A., Isenberg, E., & Kopa, E. (2013). *Does tracking of students bias value-added estimates for teachers?* (Working Paper 15). Princeton, NJ: Mathematica Policy Research.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1992). Contextual effects on the self-perceived efficacy of high school teachers. *Sociology of Education*, *65*(2), 150–167.

Reckase, M. D., & Martineau, J. A. (2015). The evaluation of teachers and schools using the educator response function (ERF). In R. W. Lissitz & H. Jiao (Eds.), *Value added modeling and growth modeling with particular application to teacher and school effectiveness* (pp. 219–235). Charlotte, NC: Information Age Publishing.

Resch, A., & Isenberg, E. (2014). *How do test scores at the floor and ceiling affect value-added estimates?* (Working Paper 33). Princeton, NJ: Mathematica Policy Research.

Rice, J. K. (1999). The impact of class size on instructional strategies and the use of time in high school mathematics and science courses. *Educational Evaluation and Policy Analysis*, *21*(2), 399–414.

Rice, J. K. (2013). Learning from experience? Evidence on the impact and distribution of teacher experience and the implications for teacher policy. *Education Finance and Policy*, *8*(3), 332–348.

Riley, T. (2014). Boys are like puppies, girls aim to please: How teachers' gender stereotypes may influence student placement decisions and classroom teaching. *Alberta Journal of Educational Research*, *60*(1), 1–21.

Roberson, L., Galvin, B. M., & Charles, A. C. (2007). When group identities matter: Bias in performance appraisal. *Academy of Management Annals*, *1*(1), 617–650.

Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Education Psychology*, *76*(6), 1000–1027.

Ross, S. M., Nunnery, J. A., Goldfeder, E., McDonald, A., Rachor, R., Hornbeck, M., & Fleischman, S. (2004). Using school reform models to improve reading achievement: A longitudinal study of direct instruction and success for all in an urban district. *Journal of Education for Students Placed at Risk*, *9*(4), 357–388.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2010). *Value added of teachers in high-poverty schools and lower-poverty schools* (Working Paper 52). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Slavin, R. E., & Madden, N. A. (1987, April). *Effective classroom programs for students at risk*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Slocum, T. A. (2004). Direct instruction: The big ideas. In D. J. Moran & R. W. Malott (Eds.), *Evidence-based educational methods* (pp. 81–94). San Diego, CA: Elsevier Academic Press.

Smith, S. P., Nelson, M. M., Trygstad, P. J., & Banilower, E. R. (2013). *Unequal distribution of resources for K-12 science instruction: Data from the 2012 national survey of science and mathematics education*. Chapel Hill, NC: Horizon Research.

Sorhagen, N. S. (2013). Early teacher expectations disproportionately affect poor children's high school performance. *Journal of Educational Psychology*, *105*(2), 465–477.

Steinberg, M., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*. Advance online publication.

Stodolsky, S. S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher*, *13*(9), 11–18.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, *23*, 565–578.

Walsh, E., & Lipscomb, S. (2013). *Classroom observations from Phase 2 of the Pennsylvania Teacher Evaluation Pilot: Assessing internal consistency, score variation, and relationships with value added. Final report*. Cambridge, MA: Mathematica Policy Research.

Warburton, E., & Torff, B. (2005). The effect of perceived learner advantages on teachers' beliefs about critical-thinking activities. *Journal of Teacher Education*, *56*(1), 24–33.

What Works Clearinghouse. (2007). *Corrective reading. What Works Clearinghouse intervention report*. Rockville, MD: Author. Retrieved from http://files.eric.ed.gov/fulltext/ED497718.pdf

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology*, *35*(3), 521–551.

Whitehurst, G. J. (2015). Getting classroom observation right: Lessons on how form four pioneering districts. *Education Next*, Winter, 63–68.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brookings Institution.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81.

Wolfe, E. W. (2004). Identifying rate effects using latent trait models. *Psychological Science*, *46*(1), 35–51.

Woolf, S. B. (2015). Special education professional standards: How important are they in the context of teacher performance evaluation? *Teacher Education and Special Education*, *38*(4), 276–290.

Xu, Z., Özek, U., & Corritore, M. (2012). *Portability of teacher effectiveness across school settings* (Working Paper 77). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Xu, Z., Özek, U., & Hansen, M. (2015). Teacher performance trajectories in high- and lower-poverty schools. *Educational Evaluation and Policy Analysis*, *37*(4), 458–477.

Yoshikawa, H., Aber, L. J., & Beardslee, W. R. (2012). The effects of poverty on the mental, emotional, and behavioral health of children and youth: Implications for prevention. *American Psychologist*, *67*(4), 272–284.

Zohar, A., Degani, A., & Vaaknin, E. (2001). Teachers' beliefs about low-achieving students and higher order thinking. *Teaching and Teacher Education*, *17*(4), 469–485.

Author

ANTHONY MILANOWSKI is a senior researcher at Westat, an employee-owned research corporation in Rockville, Maryland; AnthonyMilanowski@westat.com. He is currently providing technical assistance to the U.S. Department of Education's Teacher Incentive Fund grantees on educator performance evaluation, compensation, and human capital management. Before joining Westat, he was an assistant scientist with the Wisconsin Center for Education Research at the University of Wisconsin-Madison, where he conducted research on educator performance evaluation and compensation.