Abstract: 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 bias in classroom observation ratings as well. As a remedy, adjusting observation ratings for the relevant characteristics of teachers’ classrooms has been proposed as a remedy, analogously to how value-added teacher effectiveness estimates are developed. After a brief discussion of the potential disadvantages of adjustment, this paper distinguishes 5 potential reasons for differences in practice evaluation ratings between teachers of more or less disadvantaged and/or higher and lower achieving students. It argues that whether the relationship between teacher practice ratings and student characteristics represents bias or unfairness, and whether adjustment is appropriate, depend both on the actual causes of the relationship and the use to be made of the ratings. The paper uses analogies with the employment testing literature to place the differences in context and to suggest ways to identify their potential causes.

Prepared for the 41st Annual Conference of the Association for Education Finance and Policy, Denver, CO. Session 1.04 March 18, 2016
Lower Practice Ratings for Teachers of Disadvantaged Students: Bias or Reflection of Reality?  
(Or Just Murky Waters?)

Introduction

Value-added estimates of teachers’ contributions to student achievement have been criticized for bias relating to the 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. Perhaps in response, proponents of value-added have not been long in pointing out that measures of teaching practice used in evaluation, typically based on classroom observations, may be subject to similar types of bias (Cohen & Goldhaber, 2016). This bias could actually be worse, 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. Whitehurst et al (2014), after finding that teacher rankings based on observation scores were associated with students’ level of prior achievement, concluded that:

“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)

Chaplin et al. (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 multi-measure performance evaluation system. Lazarev and Newman (2015), using data from the Measuring Effective Teaching (MET) Project, also found positive correlations between practice scores based on videos of classrooms and classroom average pretest scores. Way back in 2005, Borman and Kimball 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. Steinberg and Garret (2016) have recently taken this line of research several steps 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’ instructional skill.

These findings have led to suggestions that such ratings be adjusted for student poverty or prior achievement analogously to how value-added estimates are typically estimated with controls for pre-existing learning and student demographic characteristics. However, there are some reasons to be cautious about adjustment. Despite their findings, Lazarev and Newman (2015) wrote that:

As a policy suggestion, however, such an adjustment may be inappropriate if teacher assignment is not random. 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. (p2.).

Research stimulated by issues with value-added teacher effectiveness measures has produced evidence that teacher assignment is often not random (e.g., Dietrle et al, 2015; Kalogrides et al, 2011; Kraemer et al, 2012). There is also fairly reasonable evidence that teachers of some disadvantaged students do experience different teaching. For example, a large scale study by Pianta et al (2007) concluded that higher SES children were more likely to be in classrooms with more positive emotional and instructional climates, though the effects were small. Raudenbush (2013) cited studies showing that teachers adjust the content and pacing of their instruction based on the average prior achievement of their students.

This paper argues that a decision to adjust scores is premature without a more in-depth consideration of why the score differences exist. As argued below, there are several potential causes of the observed relationships between student characteristics and ratings, and for some, adjustment is not the appropriate remedy. Further, adjustment could make it more difficult to assess whether disadvantaged students have less access to good teaching. Not only would real differences between teachers be obscured, but the use of evaluation ratings to improve teaching in individual cases would be undermined by removing the direct link between the rating and the description of better teaching in the rubric. Unlike value-added estimates, rubric ratings are intended to be criterion referenced, and adjustment weakens, if not breaks, the reference.
While teachers could be given two scores (an unadjusted one for formative and an adjusted one for summative uses or stakes) this risks both confusing teachers and lowering the potential motivational force.

The potential disadvantages of adjustment suggest that it is important to consider all factors that contribute to differences in evaluation ratings. Research on test fairness has long been concerned with score differences like these, and there is consensus that in many cases they are due to actual differences in group characteristics on the constructs being measured. Thus the Standards for Educational and Psychological Testing (AERA etc. 1999, p. 75) cautioned that:

“… differences in testing outcomes should in many cases trigger heightened scrutiny for possible sources of test bias, outcome differences across groups do not in themselves indicate that a testing application is biased or unfair”.

To label the instruments as biased without ruling out other causes could have negative consequences until more is known about this issue.

Better understanding of the reasons for the observed relationships between practice evaluation ratings and student characteristics starts with identifying potential causes. As a start, five are discussed here:

1. Rater bias, which includes factors that systematically influence raters to rate teachers of disadvantaged students lower than teachers of more advantaged students, even when the teachers are teaching in ways that would merit a higher rating.
2. Rubric bias, which exists when rubric being used for evaluation is based on teaching behaviors that are not effective with disadvantaged students.
3. Differences in teacher expectancy, such that teachers of disadvantaged students are more likely believe that their efforts will not lead to their students learning, and thus they reduce or re-direct effort.
4. Differences in teacher skill, such as would be found if less skilled teachers are assigned to more disadvantaged students.
5. Differences in difficulty of teaching different groups of students. If it is more difficult to teach poor or lower prior achievement students than their peers, it may not be possible without extraordinary effort or skill for teachers of these students to teach in a way that earns high ratings.

The next section elaborates on each potential cause and reviews evidence for their operation.
Potential Causes of Relationships between Teacher Practice Ratings and Student Characteristics

Rater Bias

Considerable research on race/gender bias in performance appraisal, though focusing on the characteristics of the ratee, does suggest that there could be influences of stereotype: incongruence with expectations based on the stereotype (Roberson et al, 2007). While I was unable to locate any research on bias in teacher rating based on the types of students taught, there is research that suggests that student characteristics can bias teachers’ perceptions of students (e.g., labeling of students as exceptional, race, socio-economic status). If so, it is possible that observers of teaching could also be biased as to what they observe students doing. That is, 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 low SES students, or students with lower prior levels of achievement, 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 Danielson Framework for Teaching (FFT). Coupled with the commonly-found tendency for confirmation bias in data collection and judgment (Nickerson, 1998) this could lead to situations in which, if a rater believed the students were not prepared to use higher-order thinking skills, or are likely to be less disciplined, he or she may overlook instances of teaching that disconfirm that assumption and focus on instances that are consistent with it. 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. As Catherine McClellan, one of the many who worked on the MET study wrote me:

“there is a pretty strong research base suggesting that teachers are assigned to classes of students for reasons that include reward/punishment, this may also have a component that tracks straight back to personal knowledge. The principals, who are involved in assigning the classes of students in the first place, give the *best* classes of students to the teachers they already perceive as their *best* teachers, and their observations scores confirm their pre-existing opinions” (C. McClellan, personal communication, 2015).
It could of course be counter-argued that rater bias could also be disguising an even stronger positive relationship between student disadvantage and practice ratings. In schools with higher proportions of low performing or disadvantaged students, the raters themselves could be more lenient. They may tend to make excuses for poor teaching, or may be reluctant to discourage teachers from staying by giving low ratings, in the fear that the teacher will be hard to replace. It is also possible that raters in these schools have never seen really good teaching, and are therefore likely to assume that the teaching they see is at least proficient.

Clearly, rater bias needs to be eliminated or minimized, since it could lead to conclusions that teachers are teaching worse (or better) than they are, making it difficult to improve practice or equity. However, this type of bias would likely be better addressed by rater training, rater accountability, and the use of multiple raters than by adjusting observation scores, unless one could identify specific raters as being biased. Adjustment based on student characteristics could easily risk over-adjusting that disguised actual poor teaching.

**Rubric Bias**

As mentioned above, rubric bias exists when the rubric being used for evaluation is based on teaching behaviors that are not effective with certain groups of students. This could be because poor students or students with lower prior achievement do better when techniques are used that are described at lower levels of the rubric. For example, earlier this century some reformers advocated using more structured, perhaps even scripted lessons for such students. Teachers and others have argued that it would be difficult to demonstrate behavior described at the highest level of a rubric such as the Framework for Teaching (FFT) when following the dictates of these programs. A rubric like the FFT typically defines higher practice levels as more “constructivist”, and such structured programs were often criticized for not embodying constructivist pedagogy. Another possibility is that teachers of such students have worked out their own pedagogical approaches that are not recognized by the higher levels of the rubrics typically used for teacher evaluation. Because these methods seem to work for them, they persist in using them despite the potential for receiving lower ratings.

Note that this sort of bias seems at first somewhat unlikely, because we tend to believe and expect teachers to believe that constructivist pedagogy works for all students. However, while the efficacy of constructivism draws plausibility from research on how people learn (e.g.,
Bransford et al, 1999, 2000), and rubric developers cite a research base for such techniques (e.g., Danielson, 1996) it has not yet been demonstrated that such teaching, as represented in popular evaluation rubrics, works for all students. Note also that this is the type of bias, if present, would call for more than adjusting evaluation ratings for student SES or prior achievement. It would require a basic rethinking of the rubrics or the use of a different rubric for evaluating teachers of disadvantaged students. Evaluation of special educators is an analogous and illustrative case. These teachers have argued that standard rubrics do not recognize their specific pedagogy, and some districts do in fact modify their rubrics to apply to these teachers.

**Teacher Beliefs and Expectancy**

Teachers assigned to classrooms with disproportionately poor students, or students with lower prior achievement, might well believe that they either lack the skills, or these students lack the prerequisite preparation, to use the pedagogy described in the higher levels of the rubric. Or they may falsely 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. While this could be seen as a variation of differential teaching skills, it is worthwhile to distinguish low expectations because of the research (e.g., Sorhagen, 2013; Figlio, 2005) and advocacy (e.g., Marchitello and 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 a rubric like the FFT are effective with lower performing students. Zohar et al (2001) found that many teachers believed that teaching activities based on higher level thinking skills were ineffective for low achieving students. Warburton & Torff (2005) found that teachers deemed it appropriate that less advantaged students receive fewer high critical thinking activities than more advantaged students. Wigfield & 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 can they are more effective when working with higher-ability students.
Steinberg and Garrett’s (2016) results can be interpreted to be consistent with expectancy influences on ratings in that the performance dimensions the authors interpret as requiring substantial collaboration between teachers and students more consistently show larger effects because higher achieving students are more likely to be engaged in the learning process. Student engagement is likely to depend at least in part of teachers’ attitudes. 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.

If teachers of poor or low prior achievement students can be convinced that they can succeed in applying techniques favored by the rubrics, coaching or other interventions aimed at changing teachers’ attitudes would appear to be a better approach than adjusting scores. Adjustment in this case would not only cover up the effect of low expectations, but could also reinforce teachers’ beliefs that poor or low achieving students can’t be taught in ways that accord better with higher rubric ratings.

**True Differences in Teacher Skill or Ability**

If teachers of disadvantaged/low prior achievement students actually have lower skills, we would expect that at least some of the relationship between evaluation ratings and student characteristics would be due to this skill gap. There is a substantial amount of research that suggests that at least some of the relationship between evaluation ratings and student demographics is due to lower levels of skill of the teachers assigned these students.

First, 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., Rice, 2014; Chingos and Peterson, 2011). There is also evidence that such 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., Jacob and Walsh, 2011; Harris and Sass, 2009; Milanowski, Kimball, & White, 2004) though like the relationship between value-added and experience, this relationship is strongest in the early tears 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. Since this relationship apparently also holds with value-added replacing evaluation ratings, it is reasonable to
hypothesize that those teach poorer or lower achieving students have lower skill levels. That less skilled teachers are assigned to students with lower prior achievement is supported by analyses by Steinberg & Garrett (2016), who found that ELA teachers with apparently lower levels of skill were assigned students with lower prior achievement.

**Differences in Teaching Difficulty**

If it is more difficult to teach disadvantaged students or students with lower prior achievement than their more advantaged or higher-achieving peers, then teachers of equal ability or skill would be likely to perform more poorly, on average, than teachers assigned to more advantaged or higher prior achievement students.

Common sense suggests that poor and lower prior achievement students are more difficult to teach. This seems so obvious that Berliner (2014) used an anecdote about a teacher who told him that though she didn’t change her teaching much, her evaluations changed markedly with her classroom demographics to begin his critique of using value-added for teacher evaluation. But it is not that easy to find research that definitively shows that all else equal poor or lower achieving students are harder to teach, though there are studies that provide evidence that poor students have often have emotional, social, and cognitive challenges that they must overcome to learn as quickly as more advantaged peers.

That poor students are likely to learn less appears to be supported by value-added results: 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. Newton et al, 2010; Ballou, Sanders, and Wright, 2004). However, Sass et al (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 et al (2012) found that switching between schools with substantially different performance or poverty levels does not reduce teacher performance. Xu et al (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 et al, 2013), but it is unclear whether this is due to test ceilings or similar artifacts (e.g., 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 student characteristics using MET data. He found no relationship between changes in observation-based practice ratings and changes in classroom composition. However, Steinberg & Garrett (2016) also using MET study data, but a different form of analysis, not only that found that ELA teachers with apparently lower levels of skill were assigned students with lower prior achievement, but also that “[ELA, 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 seems like good evidence that not only do teachers change the way they teach when classroom composition changes, but that it is more difficult to teach in ways promoted by the rubrics when assigned a classroom with lower average prior student achievement. However, it is puzzling that the effect is more apparent for English language arts teaching compared to mathematics, especially since the rubrics used were not subject-specific.

If teaching poorer or lower prior achievement students is more difficult for most teachers, and that this was the primary cause of systematic rating differences, most would agree there is a fairness issue. If teachers of roughly equal skill do not get the same opportunity to receive a high rating due to the student they are assigned to teach, this would be like assigning a harder test to some students then to others. Adjustment would seem appropriate just as test equating adjusts scores on different test forms that have different total difficulties. Assignment of teachers to classrooms with different proportions of difficult-to-teach students is like giving tests that are not parallel and using the results with one set of cut scores.

Again, however, it is important to recognize that it is not the evaluation instrument that is biased: teacher assignment is biasing the results, when they are interpreted as the relative ability of teachers to enact and facilitate desired behaviors. If the goal is to assess a teacher’s “true” skill relative to other teachers, this situation is unfair. It is also inefficient, because we would be more
likely to spend professional development dollars or “deselect” teachers who are disproportionately assigned these students, though these teachers may not be worse than average. Or we may reward teachers who are no better. Adjustment in this case is seems justified. But if the results were used only for formative purposes, that conclusion seems less appropriate. If comparison among teachers is not the issue, but comparison to a standard of desired practice, the solution would be more support for these teachers. It would also seem appropriate to track the non-adjusted ratings to assess the efficacy of support interventions.

To Adjust or Not to Adjust?

There appears to be relatively little evidence about the strength of the causes of differences in ratings discussed above. Both rater bias and rubric bias are plausible, but there is little direct evidence that they exist. There is reasonable evidence that differences in teacher expectancy and skills could be operating, and mixed evidence for whether poor or lower achieving students are harder to teach.

These potential reasons for lower ratings of teachers who teach poor or lower-achieving students can be organized into two broad classes: situations in which teachers of these students are in fact teaching as intended by the designers of the evaluation systems, but that this is not recognized, most likely because of rater error, and situations in which these teachers are not teaching as intended, whether because they don’t believe it is possible, they don’t have the skills, they try, but fail because doing so with their students is too difficult, or because they (rightly) believe it doesn’t work. Whether adjustment of scores for student characteristics is appropriate depends on which of these factors are operating, and their relative importance. Table 1 below summarizes the argument about if and when to adjust.1

---

1 Lazarev and Newman (2015) also found that grade level was correlated with evaluation ratings, and an interaction between grade level and prior achievement. As they observe, this further complicates adjusting ratings.
### Table 1: Sources of Rating Differences and Appropriateness of Adjustment

<table>
<thead>
<tr>
<th>Teachers teaching as intended?</th>
<th>Potential reason for lower ratings</th>
<th>Adjust?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Rater bias</td>
<td>Yes, if severe raters can be identified</td>
</tr>
<tr>
<td>No (true differences in teaching practice)</td>
<td>Rubric bias</td>
<td>No; change rubric</td>
</tr>
<tr>
<td></td>
<td>Teacher expectancy</td>
<td>No; provide training, change assignment practices</td>
</tr>
<tr>
<td></td>
<td>Teachers assigned to these students lack the skills</td>
<td>No; provide training, change assignment practices</td>
</tr>
<tr>
<td></td>
<td>Poor/lower achieving students are more difficult to teach</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Of course, if multiple causes are at work, it might be quite difficult to decide which is dominant. Does this argue for holding teachers harmless by possibly “overadjusting”, rather than risk unfairness, or not adjusting at all? This topic will be taken up after considering methods for assessing whether each of these potential causes of lower ratings are operating.

**Analyses for Exploring Bias Sources**

This section describes analyses school districts or other jurisdictions observing rating differences for teachers of different groups of students might want to perform to assess the plausibility of the potential reasons discussed above. Given the multiplicity of potential causes, multiple methods will likely be necessary to substantiate a conclusion of bias. Multiple lines of evidence are needed to support a claim of bias just as multiple lines of evidence are needed to support a claim of validity (e.g., Herman et al, 2011).

**Rater Bias**

The obvious way to examine rater bias is to have another set of raters rate the teachers’ behavior. Those raters would, of course, have to be extensively trained to avoid bias, and would have to observe independently. Ratings of video recorded teaching occasions would make this easiest, but one would have to establish the similarity of ratings made of videos with ratings made from live observations.
It might also be useful to have an independent set of raters look at the classroom artifacts and other documents that are used to assess those performance dimensions that do not depend on classroom observations. It is likely that evaluators who do not know much about the prior achievement or socio-economic status of the teachers’ students would find fewer potentially biasing cues in that evidence, and thus, if the normal raters were biased, would rate these dimensions higher than the normal raters.

Multiple raters would also allow examining the severity/leniency of individual raters. If more severe raters tended to be concentrated in schools where classrooms are poorer or prior student achievement lower, this could contribute to differences in ratings among teachers with different proportions of poor or lower-achieving students. Rater adjustment techniques might then be employed, and then one could assess whether adjusting for severity/leniency reduces the correlation of ratings with student disadvantage or prior achievement. Such analyses require that raters be totally not confounded with classroom characteristics, however. Each rater would have to rate at least some teachers of both high and low disadvantaged students. Ideally all raters would be compared to all others on teachers of varying proportion of disadvantaged students.

Even where multiple raters are not available, it could be instructive to examine whether the relationship between classroom characteristics and ratings varies across raters. One might expect that not all raters would be biased, and that for some, the relationship between classroom poverty or prior achievement and ratings would be minimal. If ratings by the large majority of raters show similar relationships with student characteristics, this would be more consistent with an interpretation that teachers of poorer or lower achieving students were actually less likely to teach as valued by the evaluation rubrics. The usefulness of this analysis depends on the extent to which the classroom characteristics of interest vary among the teachers rated by each rater, however. In districts with lots of schools where all classrooms have similar proportions of poor students, separating rater from classroom effects will be more difficult.

In districts without multiple raters, student survey results could also be used to assess the potential for rater bias, at least for those performance dimensions that correspond to constructs measured by the surveys. Ferguson and Danielson (2015) provided some evidence that the Framework for Teaching and the Tripod student survey both cover some similar dimensions of teacher performance. If rater bias is a strong cause of the differences in ratings, we would then
expect that student survey responses would not be as strongly associated with student characteristics as the ratings are. Chaplin et al (2014) provide an example. They reported that the correlation of the composite rating for teachers was correlated with student poverty at -.15, while the Tripod composite score was correlated -.09. At the dimension (component) level, all of the evaluation ratings were negatively correlated with student poverty (between -.14 and -.18) but the correlations of the Tripod dimensions (except in one case) with student poverty were lower. Of interest is that the strongest negative correlation (-.26) is between survey scores on the “controls” component (involving minimize misbehavior, promoting good behavior, and the class staying busy and on task), which in terms of content is most related to the Framework’s Classroom Environment domain. The correlation between ratings on that domain and student poverty was -.15. These results suggest the possibility of some rater bias, but also suggest that at least some of the ratings difference associated with poverty are not simply due to rater bias.

Another approach to bias exploration would be to manipulate aspects of the rating situation as part of rater training and certification. Numerous studies of student ratings of post-secondary teaching have been done in which the characteristics of the teacher (e.g., ethnicity) were manipulated. Simulations in which the same basic performance was exhibited by ratees with different characteristics are common in rater bias research (e.g., Feldman and Hilterman, 1977). While it would be more difficult to manipulate student characteristics to look at their effects on teacher ratings, it might be possible to have actors (teachers and students) follow scripts portraying different teaching behaviors with different ethnic mixes and different cues as to the poverty and achievement status of the school or classroom. This approach would allow the identification and retraining of potentially biased raters.

Rubric Bias

If a bias free criterion of performance existed, then it would be quite easy to determine if the rubrics themselves, when applied faithfully by calibrated raters, were privileging teaching behaviors that were not associated with student learning but were associated with student disadvantage. To some extent, value-added estimates of teacher effectiveness could be considered relatively bias free, at least in that they control for students’ prior achievement and in most cases student poverty. Thus they have already been purged of much if not all of the effects of the two characteristics we might consider using to adjust practice ratings.
In employee selection research, differential prediction is a standard method for testing for bias against protected groups. This involves regressing criterion scores on predictor (selection test) scores, an indicator for group membership, and the interaction of predictor score with group membership. Evidence for bias exists if (when appropriate assumptions are justified), the coefficient for the interaction is statistically significant (Aguinas, 2004).

In this case, the criterion would be the teacher’s value-added estimate, and the predictor, the teacher’s practice rating. Instead of a group membership indicator, here the average subject pre-test score or proportion of poor students taught would be used, along with the interaction between that covariate and the practice score. If the rubric were biased against teachers of low performing or poor students due to not recognizing the specific teaching practices that are differentially effective with these groups, we would expect a positive coefficient for the average prior achievement-practice rating interaction or a negative coefficient for the proportion of poverty-practice rating interaction (depending on our focus). The interaction coefficient would suggest that the relationship between the practice rating and value-added would be weaker for teachers with higher proportions of poor students or with students with lower average pre-test scores, because the practice rating fails to capture the practices that contribute to better learning for these students. As illustrated in Figure 1 below, practice ratings would underpredict the effects of teachers who are effective with these students and overpredict the effects of teachers who are less effective with these students.

Figure 1: Hypothetical Relationship Between Value-Added and Teacher Evaluation Ratings for High, Average, and Low Poverty Classrooms
When the rubrics don’t work the same for teachers of high versus low poverty students, the strength of the relationship between ratings and value-added is different for the two groups, with the teaching behavior described by the rubric having a much higher return, in terms of value-added student achievement, for teachers of low poverty students.

Rubric Bias and Differential Item Functioning

Differential item functioning is the name given to situations in which a subset of test items show differences in scores (e.g., percent correct) between examinees with the same overall ability level. Translated into performance evaluation terms, ratings on a specific dimension or component\(^2\) of a performance evaluation rubric might differ among teachers with different proportions of disadvantaged students, though these teachers’ overall performance, as measured by total evaluation scores (i.e., the sum or average of all other dimension ratings) are the same. In other words, it is not necessarily that the entire rubric is biased, but only specific parts or dimensions.

Were teachers with higher proportions of disadvantaged students rated lower on particular performance dimensions, even where they have similar overall scores, this would be consistent with an interpretation that teachers are not teaching in this manner not because they do not have skills, but because they believe that students can’t be taught in this way or that it is more difficult to do so.

Unfortunately, characteristics of performance evaluation ratings complicate applying the kinds of differential item functioning analyses used for student and similar assessments. With standardized, multiple choice knowledge tests, the assumption is that relatively few items are biased while most others are not, and that the others are sufficient to construct a reasonably unidimensional measure of performance that is unbiased. However, for performance assessments like teacher evaluations, there are typically relatively few dimensions, and the dimensions themselves, while inter-related, are typically designed to assess different aspects of

\(^2\) A dimension or component is the most specific aspect of behavior that can be rated using a given evaluation rubric or instrument. For example, in the original Framework for Teaching, there were 22 components that could be rated ranging from student questioning to arrangement of physical space.
teaching. So it is difficult if not impossible to construct a unidimensional, unbiased performance measure out of the remaining components.

An alternative for assessing differential item functioning in performance assessments that can be applied to performance evaluation ratings is to find an external criterion representing true performance that is not expected to be biased (Penfield and Lam, 2000). Again, value-added estimates of teacher effectiveness can be called into play, as long as we are aware that the conceptual and empirical relationship will not be as strong, since value-added is not measuring teaching practice or the ability to teach according to the conceptions underlying the rubrics.\(^3\)

The analysis would focus on whether, at the same level of value-added, teachers of poorer or lower achieving students tend to be rated lower on specific teaching practice dimensions that we suspect might be biased. For example, we might compare ratings on dimensions like managing classroom behavior versus providing feedback to students or choice of instructional goals. This could help pin down what dimensions of the rubric might be biased, in the sense of promoting pedagogy that is not as effective with poor or low prior achievement students. In this case, the analysis would be based on a model like:

\[
\text{Rating}_i = \beta_0 + \beta_1 \text{classroom value-added} + \beta_2 \text{dimension rating} + \beta_3 \text{proportion of student poverty} + \beta_4 \text{rating/poverty interaction} + \varepsilon_i
\]

In the case of proportion poor students, the coefficient for percent poor would be negative, showing at each level of value-added, teacher evaluation ratings are lower for teachers with a greater proportion of poor students. A result such as the hypothetical relationship shown in Figure 2 below would provide evidence for the need to further examine the dimension in question.

\(^3\) Another complexity is value-added tends to differ by subject, while for many elementary teachers, practice ratings are made based on observations of lessons covering different subjects. If ratings are based on observations of multiple subjects, it could be more appropriate to use a composite value-added measure (e.g., reading and math) for such an analysis.
It is possible that some dimensions represent teaching practices which are not as effective for some types of students, while others represent practices that are not differentially effective. One possible scenario is that for a classroom dominated by low prior achievement students, a teacher could emphasize lower cognitive level skills and still have reasonably high value-added if the tests measuring achievement are sensitive to learning basic content that the students had not yet mastered. If ratings on dimensions more closely related to promoting these skills how a stronger relationship with student characteristics, this would suggest that the rubrics do not represent a vision of teaching that is effective for all students.

Empirically, it is not clear that differential functioning of this sort will be found. Chaplin et al (2014) found that the correlation between classroom poverty and race/ethnicity and ratings on different domains of the Pittsburgh practice rubric differed little, while Lazerov and Newman (2015) found that correlations between ratings of observed dimensions of the Framework for Teaching and average prior student prior achievement in the MET study ranged from .19 to .3, suggesting some possibility of differential functioning.

**Differences in Teacher Expectancy and Skills versus Differences in Teaching Difficulty**
Since it is fairly well accepted, based on value-added studies, that teachers in their first few years of experience are less productive, and common for less experienced teachers to teach disadvantaged students, it would seem reasonable to assess the relationship between teacher experience, teacher practice ratings, and student characteristics. If less experienced teachers have not yet developed the skills to teach in highly-rated ways, and are more likely to be assigned poorer or lower achieving students, the relationship between evaluation scores and student characteristics may be due at least in part to teacher experience. This would be easy to examine by simply controlling for experience when assessing the relationship between performance ratings and student poverty or prior achievement. If the difference in ratings largely disappears, ratings are less likely to be biased. However, one could still argue that holding less experienced teachers to the same performance standard as more experienced is not quite fair, but the unfairness is in the assignment of students to teachers rather than bias in the ratings.

An analysis like that of Steinburg and Garrett (2016) can provide evidence for the likely contribution of persistent teacher skill differences versus differences in teaching difficulty. Their approach does require that the proportions of poor or low-achieving students assigned to teachers change over time, something that might not happen in districts with low interschool mobility and homogeneous schools. Districts may have to use more years of rating data to capture enough variation to separate persistent teacher ability from transient effects of classroom composition.

Even as sophisticated an analysis as Steinburg and Garrett’s cannot, however, differentiate between teacher expectancy and differences in the actual difficulty of teaching poor or lower prior achievement students. Doing so is likely to be difficult using evaluation results and other performance measures alone. What may be required is in-depth qualitative studies that involve long term teacher observation and interviews with teachers. Assuming that expectancy and ability to teach these students vary across teachers, it could be of interest to observe and interview 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. By plotting ratings by classroom proportion of poor students, as in Figure 3 below, one can identify teachers to study. In Figure 3 below, if teacher expectancy was an important treason for rating differences, we would expect to hear and see more indicative of low expectations, while moving downward on
the right side of the average proportion of classroom poverty line from B to D. We 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.

Figure 3. Plot of Hypothetical Ratings by Classroom Poverty

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 poor students that do not use the techniques. If the reason for less use of the techniques associated with higher ratings is due to lower expectancy, we would expect to hear teachers not using them talk about expectations rather than skills as the rationale, and to hear teachers that do use them express confidence that the techniques can be used with their students. Of course, this analysis would require the employment of skilled observers/interviewers who both understand teaching and can gain the trust of the teachers being studied.

A plot like Figure 3 also illustrates that even if teaching difficulty is the major reason for the relationship between ratings and classroom characteristics, some teachers may not find teaching...
poor or lower-performing students as difficult as other do. These teachers may be more skilled in
general, more skilled with certain types of students, or perhaps just lucky in having personality
traits that help them hit it off with their students. Then should we conclude, for example, that
teacher D is a better teacher than teacher A, because her/his ratings are just as high but were earned
teaching more challenging students? Just as the possibility of differential value-added complicates
the use of value-added estimates to conclude that one teacher is better than another (Lockwood and
McCaffery, 2009), it may be that we need to think more deeply about what we mean by teaching
ability or performance and whether practice ratings made using current techniques capture all
that is important in teaching.  

Following Lazerev and Newman (2015) it could be appropriate to control for grade level when
examining the relationship between ratings and classroom characteristics. They found that the
relationship was larger in the highest grade they studied (8th) and least in the lowest (4th). 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. It might also be appropriate to assess the degree to which school and classroom
conditions are related to lower teacher performance ratings. It is possible that at least between
schools, there are differences in resources and materials, or even physical facilities, related to
student poverty and make teaching more or less difficult. 5 Within schools, it is possible that
teacher of lower – performing students teach larger classes where it is arguably more difficult to
engage students or employ some constructivist pedagogies.

Discussion

This paper has argued that there are several potential causes of differences in performance
evaluation ratings between teachers of disadvantaged or low prior achievement students and
teachers of more advantaged, or higher achieving students. It has further argued that whether
teacher evaluation ratings should be adjusted for student characteristics, in an attempt to be fair
to teachers, should be considered in light of which of the potential causes is most likely to be at

4 However, it could be argued that current rubrics and practices get us “close enough”, as value-added effectiveness
estimates are often argued to do. It might be worthwhile to ignore some of the issues with practice ratings just as it
could be for value-added, to obtain the benefits of their use.
5 Of interest is that Walsh and Lipscomb (2013) found that a substantial proportion of variance in practice domain
ratings was between schools in Pennsylvania. This raises the possibility that some of the differences in ratings might
be due to school effects or conditions that are correlated with student poverty or average school-wide achievement.
work. The clearest case for adjustment is where there are true differences in the difficulty of teaching different kinds of students. As argued above, 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 in general 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 is invalid.

It would seem to me that the first order of business would be to try to rule out or fix rater and rubric bias. These potential bias sources are the most pernicious, the former because it disguises what is really happening in the classroom and the latter because it distorts the desired incentive to teach in ways that benefit students. These bias sources may also be the easiest to identify.

While there are a number of strategies that could be used to assess the effects of the other potential causes, it is not likely that true difficulty and teacher expectancy can be separated. If the sources or causes cannot be separated and the dominant effects identified, would adjustment be the best course, even risking over adjustment? In my view this depends on three considerations:

1. Have rater bias and rubric bias be ruled out?
2. What is the use to be made of the ratings?
3. What is the potential for equalizing classroom compositions across teachers?

If the major use of ratings is to track teacher performance and provide feedback to improve it, then adjustment has the potential to obscure needs for improvement and backburner efforts to ensure poor or lower performing students have access to the kind of instruction reformers prefer. The experience with value-added is instructive here. While most accept the need for adjusting for poverty in value-added, the adjustment has apparently put the problem out of mind. Few if any jurisdictions track whether teachers or schools are reducing the size of the negative coefficients on poverty (or ethnicity). Adjustment of overall practice ratings would be especially problematic if rating differences are primarily due to one or two performance dimensions. If ratings differences are strongly related to teacher experience, it would seem more appropriate to hold newer teachers to lower standards, with supports for improvement, rather than disguise their need to improve skills by adjustment of ratings.
If the major use is to make between-teacher comparisons or rankings for high stakes decisions, the case for adjustment is much stronger than if the use is to identify teachers in need of help in improving instructional practice. Teachers appear to be quite sensitive to perceived unfairness in evaluation, and, especially if financial consequences stem from evaluation ratings, the current incentives to teach less poor and higher achieving students would only be strengthened.

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 student difficulty introduce unfairness, invalidate comparisons, and compromise generalizability. As Reckase and Martineau (2015) pointed out, 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 can’t be equated by equalizing classroom composition, adjustment would be justified. Their discussion suggests that adjustment be done within an IRT framework rather than simply by controlling for student characteristics.

If equalizing classroom composition across teachers a realistic alternative, this would be preferable to adjustment for most uses. This would be especially appropriate for inexperienced teachers. In some ways this is no more radical an idea that using student achievement growth to evaluate teachers. At the school level, it seems possible, though of course much less so between schools and districts. Cross-school comparisons within districts would seem to be the most critical area, because any ranking for consequences is likely to be done within the district.

The points raised in this paper also suggest that a more careful use of the term “bias” is in order. It would be useful to differentiate between bias and unfairness. The statistical definition of bias is related to the degree to which an estimated parameter, value, or score differs systematically from the true one. Test bias is defined by the Standards as a characteristic of the test not related to the construct to be measured that differentially affects the performance of different groups. Unfairness is less precisely defined, but tends to be related to undeserved advantage or disadvantage associated with the use of a test or assessment. If we want to use “bias” in a way closer to the technical definitions, then differences due to some of these causes are not bias, though they may lead to unfairness. If teachers of poor or low prior achievement students are not using techniques that are described at higher levels of evaluation rubrics while teachers of other students are, this is not bias. Bias does seem like the term for the situation in which, because of
preconceptions, raters do not recognize practice that is reflected in higher rubric levels for teachers of some kinds of students but not others. So, in place of Steinberg and Garrett’s (2016, page 20) way of posing the problem:

… when information about teacher performance does not reflect a teacher’s practice but rather the students to whom the teacher is assigned, such systems are at risk of misidentifying and mislabeling teacher performance.

It seems to me more accurate to say, if rater bias and rubric bias can be ruled out, that while the information may reflect a teacher’s practice, if that practice is substantially influenced by the students the teacher is assigned, using evaluation information to rank teachers or to represent their likely success in teaching any group of students overgeneralizes, and mislabels their relative performance.
References


