Authors: Emanuele Bardelli & Matthew Truwit.
Title: Teacher Evaluation Systems

Your full name: Emanuele Bardelli and Matthew Truwit

Affiliated authors with institutions:

Affiliation: University of Michigan

Current position: Graduate students

Title of your paper: Teacher Evaluation Systems: Measures of Instructional Effectiveness or Mechanisms of Structural Bias?

Abstract (300 words)

Since the introduction of Race to the Top a decade ago, American systems of teacher evaluation have risen considerably in prevalence; almost as rapid, however, has been the growth in the evidence base about their potential bias against teachers from historically marginalized groups. Our work seeks to contribute to this burgeoning field of literature by investigating the potential bias in evaluation systems across the spectrum of instructional effectiveness against teachers with minoritized racial and ethnic as well as gender identities. Employing data from the Measures of Effective Teaching (MET) study, we use item response theory (IRT) to assess whether teachers of different identities but comparable teaching expertise receive systematically biased observation ratings on specific items at varying levels of underlying instructional effectiveness. In preliminary analyses, we find evidence of negative bias (i.e., lower scores even after holding underlying teacher effectiveness constant) towards all teachers belonging to minoritized racial/ethnic groups in items measuring teachers' effectiveness at designing a learning-friendly environment. Moreover, these two items also have higher discrimination parameters for racially minoritized teachers than for their white peers, suggesting that evaluators are more lenient towards less instructionally effective white teachers than they are for similarly performing minoritized teachers when appraising these skills. Given the use of these measures in high-stakes decisions around educators' careers - and the different repercussions for teachers of different proficiencies - improving our understanding of structural bias in teacher evaluations can not only mitigate the direct harm done to teachers (and students) of marginalized identities but also disrupt the dominant cultural hegemony of the American educational system.

Extended summary (1000 words, excluding reference list) introduction, theoretical background, methods, preliminary findings/findings, results, reference list.

In 2009, the U.S. federal government introduced incentives for states to develop teacher evaluation systems through the Race to the Top initiative. This competitive grant program encouraged the adoption of systems of teacher accountability composed of a combination of administrator observations and value-added measures based on student test scores. Ten years later, 41 states have implemented some form of teacher evaluation system based on at least two different performance measures.

Despite their growing prevalence, researchers have found evidence of bias in these evaluation systems against teachers both of certain identities and in particular kinds of classrooms. For example, Campbell and Ronfeldt (2018) reported that expert ratings of instruction are sensitive to teachers' ethnic and racial identities (ERIs) and gender - specifically finding that male and Black teachers tended to score lower than otherwise similar colleagues. However, they also found that these differences were largely driven by differences in classroom composition, wherein teachers with higher percentages of Black, Hispanic, and male students received lower evaluations than their colleagues, even after random assignment of classes. Steinberg and Sartain (2020) identified additional elements of student background (i.e., prior achievement, socioeconomic status, and disciplinary record) that significantly influenced teacher evaluation scores. They similarly argued that variation in evaluation scores among teachers of different ERIs is largely due to bias resulting from differences in classrooms and schools, rather than real differences in performance across teachers. More recently, Campbell (2020) described how Black women receive systematically lower evaluation scores and are twice as likely as their equally instructionally effective white female colleagues to be placed on punitive improvement plans - even *after* accounting for classroom and school characteristics.

Together, these papers suggest that, while incorporating student composition can help explain some of these differences in observation ratings, bias against teachers of color may in fact be best understood as a dynamic interaction between teachers' overlapping identities and school contexts. As such, there is a critical need for work that explores the interplay of teachers' ERIs, gender identities, and teaching contexts along distinct dimensions of instructional effectiveness in order to better understand how structural biases in evaluation systems could negatively impact the career trajectories of already marginalized teachers.

Research Questions

(1) Do observers exhibit systematic bias against teachers of different identities when evaluating particular facets of instructional effectiveness? If so, on which facets and for whom?

(2) How do these biases differ across the continuum of instructional effectiveness?

(3) What role do differences in teachers' classroom, school, or district characteristics play in explaining these biases?

Data

We use data from the Measures of Effective Teaching (MET) study to further explore the relationships between teacher ERI, gender, school context, and observation ratings (OR). The MET study involved the videotaping of 3,000 teachers from seven school districts across the United States over the course of two school years. Teams of trained assessors rated the quality of instruction in these videos using a variety of observation instruments. We match these evaluation data with additional information on teachers' ERI and gender identities and on the characteristics of the classrooms, schools, and districts in which they teach.

Methods

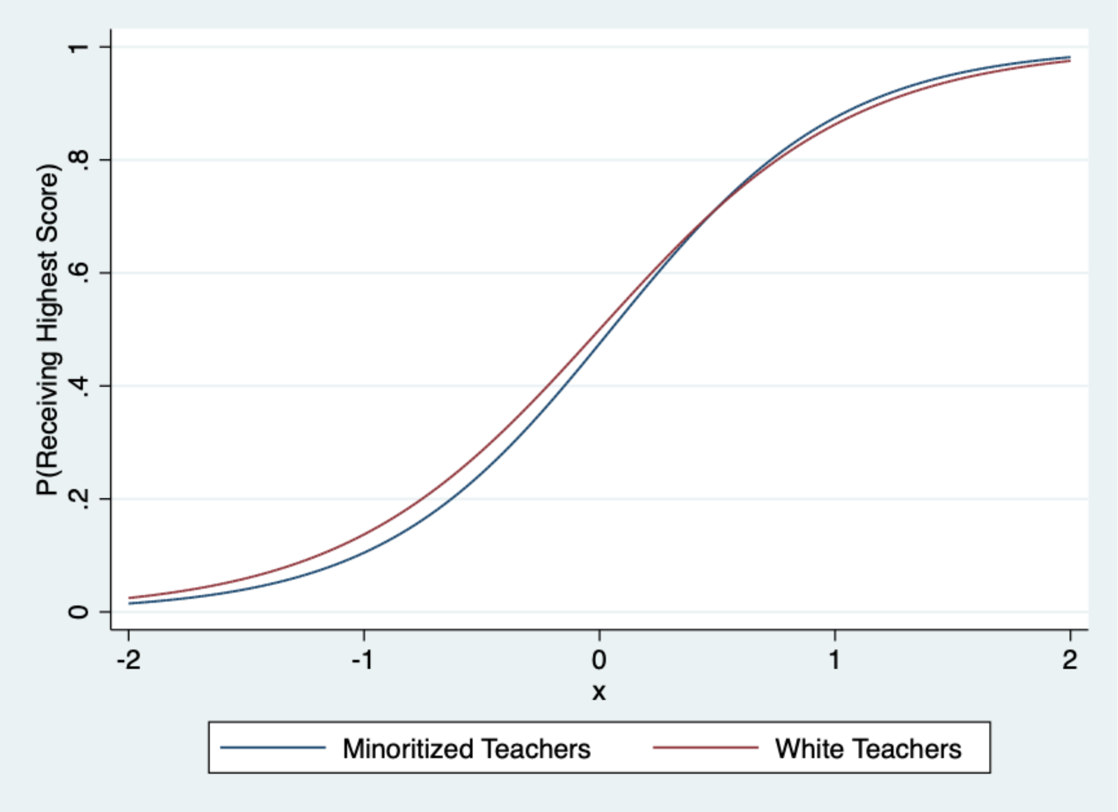
We use item response theory (IRT) to assess the extent to which individual indicators are biased against teachers belonging to particular ERI and gender identity groups. IRT offers a robust set of analytical tools to investigate for bias via assessing items for differential item functioning (DIF) and has been deemed suitable for the analysis of teacher evaluation scores (Kraft et al., 2019). Intuitively, we would find evidence of DIF when scores on an individual rubric item differ for teachers from specific intersections of ERIs and gender identities even when their underlying level of instructional effectiveness is the same (Furr & Bacharach, 2013). We examine each item for evidence of both uniform - affecting all teachers of a certain identity regardless of proficiency - and non-uniform DIF, wherein the extent of the bias varies depending on a teacher's underlying level of instructional effectiveness. Given that prior literature has shown that observation ratings appear sensitive to student and school characteristics, we also plan to estimate the extent to which any bias on specific items is explained by differences in student, school, or district characteristics, such as students' ERIs and gender identities, socio-economic status, and prior achievement.

Authors: Emanuele Bardelli & Matthew Truwit.
Title: Teacher Evaluation Systems

Findings

Although still early in our analysis, preliminary results confirm IRT as a viable approach to study DIF in observation rubric items. Exploratory analyses of one instrument have found the presence of both uniform and non-uniform DIF (see Figure 1 for an example of DIF in one item). We find evidence of negative bias (i.e., lower scores even after holding underlying teacher effectiveness constant) towards *all* teachers belonging to minoritized racial/ethnic groups in an item measuring teachers' effectiveness at designing a learning-friendly environment. Moreover, this item also has a higher discrimination parameter for racially minoritized teachers than for their white peers, suggesting that evaluators are especially lenient when appraising this skill in *less* instructionally effective white teachers compared to in similarly performing minoritized teachers. We find similar patterns for other environment-oriented facets of instructional effectiveness, including managing student behavior and developing a respectful culture.

Figure 1. Item Characteristic Curves for "Environment" by Teacher Ethnic Racial Identity



Significance

As the popularity of teacher accountability measures has grown over the past decade, so too has the evidence base of their bias. The consequences of relying on a racially prejudiced system of teacher evaluation in high-stakes decisions around educators' careers are not only of obvious direct harm to teachers (and students) of marginalized identities but also another example of the ways in which educational institutions reproduce racialized and gendered norms. We hope our findings contribute to this growing literature investigating the equity of teacher evaluation systems by deepening our understanding of the dynamic nature of their racial/ethnic and gender biases for teachers across all levels of instructional effectiveness and building momentum toward the consideration of more just and equitable alternatives for measuring teaching quality.

References

Campbell, S. L. (2020). Ratings in black and white: A quantcrit examination of race and gender in teacher evaluation reform. Race Ethnicity and Education. https://doi.org/10.1080/13613324.2020.1842345

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? American Educational Research Journal, 0002831218776216. https://doi.org/10.3102/0002831218776216

Furr, M. R., & Bacharach, V. R. (2013). Psychometrics: An introduction. SAGE.

Kraft, M. A., Papay, J. P., & Chi, O. (2019). Teacher skill development: Evidence from performance ratings by principals. In EdWorkingPaper. https://doi.org/10.26300/sad5-cz73

Steinberg, M. P., & Sartain, L. (2020). What explains the race gap in teacher performance ratings? Evidence from Chicago public schools. Educational Evaluation and Policy Analysis, 0162373720970204. https://doi.org/10.3102/0162373720970204