Exploring Race and Gender Gaps in Classroom Observation Scores in Tennessee

TN EDUCATION RESEARCH ALLIANCE

December 2021

Jason A. Grissom, Brendan Bartanen, and Ashton Toone[1]

For teacher evaluation to help schools improve, educators must be able to rely on the measures the evaluation system creates. Teacher evaluation can affect teachers and teaching via multiple mechanisms.[2] One is using classroom observations to identify areas in which teachers excel or could improve and to create opportunities for feedback and other supports. Another is to generate measures of effectiveness that can be used to inform compensation, retention, placement, and other personnel decisions. High-quality information about teacher performance is critical to both mechanisms.

In Tennessee, among the different measures that comprise the overall evaluation score—called the *level of effectiveness,* or LOE—teacher observation scores are the most important, comprising 50–70% of a teacher's final rating.[3] The significance of these scores motivated TERA researchers to analyze patterns in scores teachers receive by teacher characteristics. In particular, we focused on teacher gender and race, important factors given ongoing efforts to diversify Tennessee's teacher workforce. Research from other states has found that male and Black teachers often are rated lower than their female and White colleagues (e.g., Campbell & Ronfeldt, 2019; Steinberg & Sartain, 2020), which may make it less likely that they stay in the profession (Drake, Auletto, & Cowen, 2019).

This brief summarizes the main results of this analysis, which is reported on in more detail in Grissom and Bartanen (2021).[4]

The analysis uncovers three main findings:

1. ***Black teachers and male teachers in Tennessee consistently receive lower classroom observation scores than their White and female peers each year, across every observation system (e.g., COACH, TEAM), and at every school level.***

2. ***Black teachers and male teachers receive systematically lower observation scores than their White and female peers even when they have similar qualifications and their students achieve similar test scores and other outcomes.***

3. ***While we have few clues as to what could be driving the gender gap in observation scores, the magnitude of the race gap is influenced by several factors. These include the racial isolation of Black teachers, the differing characteristics of students who are assigned to Black and White teachers, and the race of the teacher's observer.***

How Teachers Are Assigned Observation Scores

In Tennessee, observation scores are assigned to teachers by certified evaluators—usually the school principal or an assistant principal—using one of the state's approved observation rubrics. By far, the most common rubric districts employ is the TEAM rubric, used to observe more than 80% of teachers each year. Some districts have adopted alternative approved rubrics (e.g., TIGER, COACH, and TEM), as have numerous charter management organizations. Pursuant to Tennessee State Board of Education policy, the number of times a teacher is observed for scoring purposes varies according to the teacher's licensure status and prior year performance, with lower-rated teachers and those with practitioner (rather than professional) licenses observed more often.[5]

The TEAM rubric has four domains: instruction, planning, environment, and professionalism. The first three are scored from classroom observations throughout the year. The fourth typically is scored summatively at the end of the school year. Each domain consists of multiple indicators. The rubric defines performance on a scale of 1 ("significantly below expectations") to 5 ("significantly above expectations") on each indicator. Indicator scores are averaged to produce a final observation score.

Analyzing the Observation Score Gap

DATA

We analyzed administrative data provided by the Tennessee Department of Education from all school years from 2011-12 to 2018-19. The data set included deidentified information about educators' demographic and professional characteristics, including race/ethnicity, gender, work roles held, years of experience, and highest degree earned. It also included teacher observation information. For teachers in the TEAM observation system (more than 80% of educators in the state), this information usually contained item-level scores from each observation, plus information about the rater (e.g., position, race/ethnicity, gender). For teachers in other systems, information typically was more limited, containing only final average scores, though finer-grained information was available for some systems in some years. Teacher information was merged with information about characteristics and performance of the students they taught and other information about their schools (e.g., enrollment size, demographic composition).

METHODS

The study proceeded in three stages. First, we analyzed gaps between male and female and Black and White teachers across the state and in different contexts—for example, in schools with varying student demographics. Second, we assessed whether gaps were still present among teachers with the same qualifications (e.g., experience levels and degree attainment) and job performance on other measures, such as teachers' "value added" to their students' achievement.[7] For this assessment, we employed *regression analysis,* a statistical technique for estimating the relationship between two factors, such as observation scores and teacher race, setting other factors to be equal. Third, after establishing that male-female and Black-White gaps in observation scores held even among teachers with similar qualifications and achievement growth performance, we delved into potential drivers of these gaps by testing whether each gap was explained by school context, characteristics of the students in the teacher's classroom, and other factors. Again employing regression analysis, this component of the analysis went even further, using additional model restrictions to compare scores assigned to male and female and Black and White teachers *within the same school in the same year,* and even across observations in the same year conducted by different raters. This last approach was especially helpful because it holds teacher characteristics equal to more clearly isolate patterns associated with rater characteristics, such as the rater's role in the school (e.g., principal, assistant principal) or their race or gender.

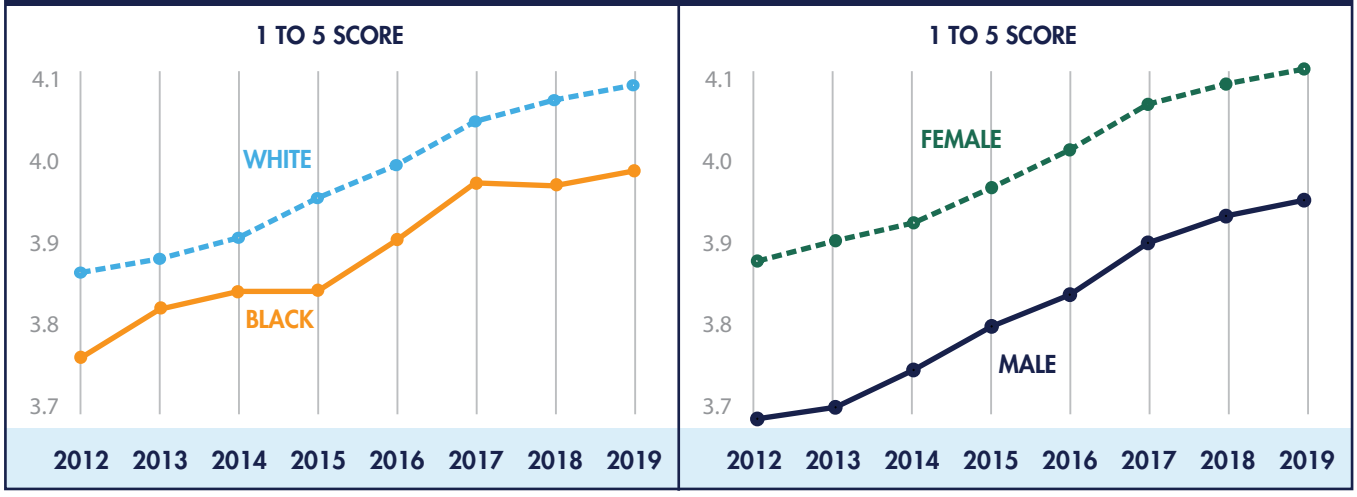
1 **Black teachers and male teachers in Tennessee consistently receive lower observation scores than their White and female peers each year, across every observation system (e.g., COACH, TEAM), and at every school level.**

Over the first eight years of the evaluation system's implementation, results show that the average observation score gap between Black and White teachers and male and female teachers has been similarly sized over time. As Figure 1 shows, even as average scores have increased each year of the evaluation system, White teachers have outscored Black teachers by about one-tenth of a rating point (on a five-point scale) each year, a statistically significant difference.

The gender gap appears even larger than the racial gap, with female teachers scoring about 0.18 points higher than male teachers over time. The gender gap is roughly additive with the racial gap, meaning that White women score, on average approximately 0.30 points higher than Black men, a gap that is roughly constant across all years of the data set.

FIGURE 1

SINCE 2012, WHITE TEACHERS AND FEMALE TEACHERS HAVE CONSISTENTLY RECEIVED HIGHER OBSERVATION SCORES THAN THEIR BLACK AND MALE PEERS.

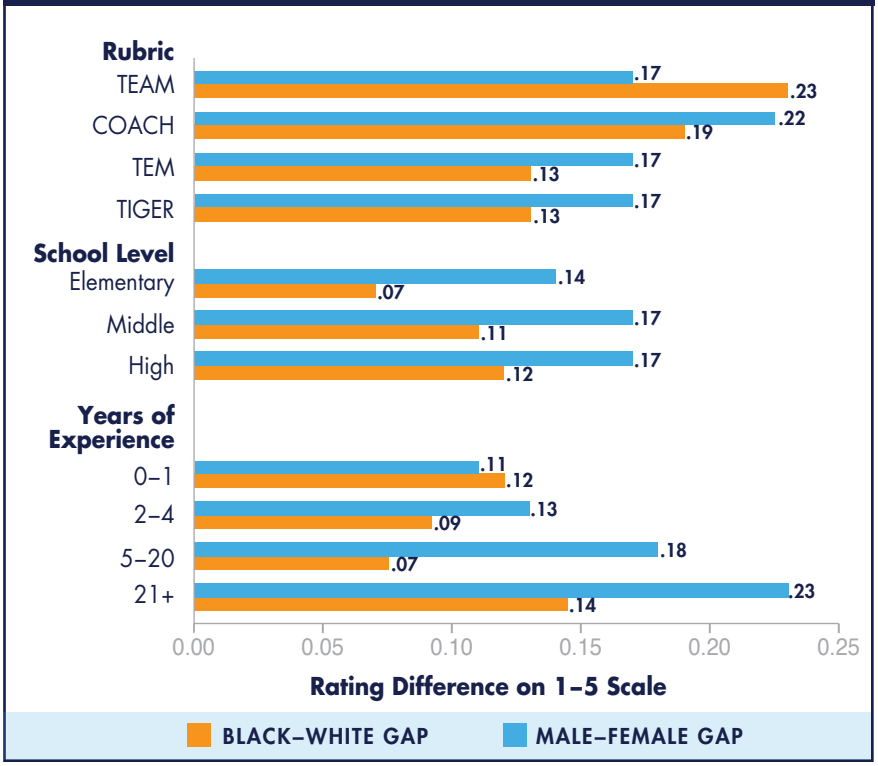


Further, as shown in Figure 2, observation score gaps between Black and White teachers and male and female teachers are present in every evaluation system used in Tennessee (TEAM, TIGER, COACH, and TEM). Racial gaps are largest in the TEAM system, the most commonly used evaluation system across Tennessee districts.[8] In TEAM, White teachers outscore Black teachers by 0.23 points, on average. Gender gaps are more consistent across evaluation systems, though the COACH system shows a slightly larger gap than the other systems.

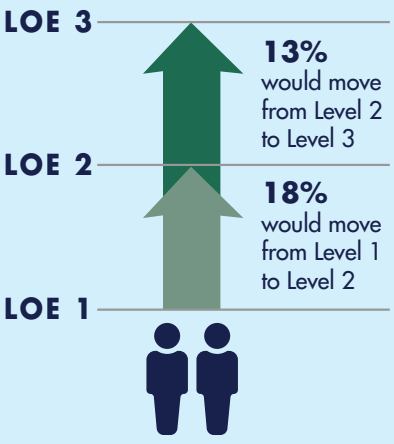
The figure also shows that gender and racial gaps exist across school levels and across different levels of teacher experience. Black and male teachers receive lower observation scores than White and female teachers across elementary, middle, and high schools. The gap is starkest in high schools, where Black teachers score roughly 0.12 points lower than their White colleagues. The gap between male and female teachers is slightly smaller in elementary schools. Turning to experience, Black teachers begin their careers receiving lower ratings, and though the gap shrinks among somewhat more experienced teachers, Figure 2 shows that it rebounds such that the gap is largest among teachers with more than 20 years of experience. By contrast, the gap increases when comparing male and female teachers with greater experience. Among teachers in their first or second year of teaching, women outscore men by 0.11 points, compared to a gap of 0.23 points among highly experienced teachers.

FIGURE 2

BLACK AND MALE TEACHERS RECEIVE LOWER OBSERVATION SCORES THAN WHITE TEACHERS ACROSS ALL EVALUATION SYSTEMS, AT EVERY SCHOOL LEVEL, AND ACROSS LEVELS OF TEACHING EXPERIENCE.



If every Black teacher had one-tenth of a rating point added to their observation score, many would receive a higher overall LOE score.



While these gaps in observation scores may seem small, they are large enough to have implications for personnel decisions. Take the average gap between Black and White teachers, which is approximately one-tenth of a rating point. **If every Black teacher had one-tenth of a rating point added to their observation score, six percent of them would have received an overall evaluation score (or LOE) that was a full point higher in a given year. Impacts would be even larger for low-scoring teachers; this addition would move 18% of Black teachers receiving an LOE of 1 to an overall score of 2, and 13% receiving an LOE of 2 to an overall score of 3.** Given the potential for personnel action for teachers identified as LOE level 1 or 2, these numbers suggest that the magnitude of these differences between Black and White teachers could have policy relevance. Additionally, the adjusted gap between male and female teachers is substantially larger, suggesting even more dramatic shifts in LOE from this hypothetical adjustment of scores.

2 Black teachers and male teachers receive systematically lower observation scores than their White and female peers even when they have similar qualifications and their students achieve similar test scores and other outcomes.

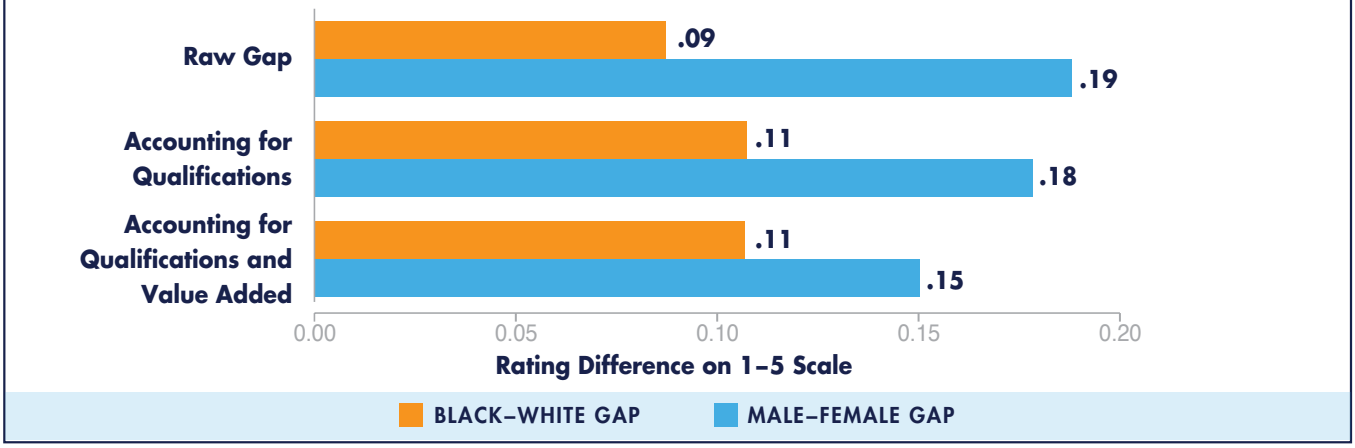
Theoretically, differences in average observation scores between Black and White teachers and male and female teachers could reflect average performance differences between these two groups. Such performance differences could arise if, for example, more experienced teachers tend to be more effective, and White teachers have higher average experience than Black teachers.

To account for this possibility, we compared observation scores for Black and White teachers and male and female teachers with the same qualifications—measured by years of experience and highest degree held—and, for teachers in tested classrooms, their estimated contribution to students' achievement growth (called "value added") in the year the observation scores were assigned. If experience, degree, or value added accounted for the difference in scores, we might expect to see the gap between Black teachers and White teachers or men and women shrink or even disappear as we add these factors to the model.

We do find that the estimated average observation score gap between male and female teachers decreases by roughly 20% when accounting for teachers' qualifications and contributions to achievement growth. Surprisingly, however, the estimated gap between Black and White teachers marginally *increases*. As Figure 3 shows, among similarly qualified Black and White teachers whose students achieve similar growth, White teachers outscore Black teachers by about 0.11 points.[9] Likewise, female teachers score approximately 0.16 points higher than similarly qualified and effective male teachers, on average.

FIGURE 3

ACCOUNTING FOR TEACHER QUALIFICATIONS AND STUDENT OUTCOME DIFFERENCES, WHITE TEACHERS STILL RECEIVE HIGHER OBSERVATION SCORES THAN BLACK TEACHERS.



These patterns raise concerns that gaps in observation scores reflect some form of bias in the evaluation system or its processes. Here, we use the word bias in a statistical sense: there appear to be systematic departures between what the observation system hopes to measure (a teacher's effectiveness) and the measures it actually produces (the observation scores). These systematic departures correlate with race and gender in ways that disadvantage Black and male teachers relative to their white and female colleagues, which, as we note above, can impact overall LOE ratings. Unfortunately, there could be multiple drivers of these systematic differences that are difficult to disentangle with existing data, though the next section describes a few analyses that aimed to shed additional light on some possibilities.

3 **While we have few clues as to what could be driving the gender gap in observation scores, the magnitude of the race gap is influenced by several factors. These include the racial isolation of Black teachers, the differing characteristics of students who are assigned to Black and White teachers, and the race of the teacher's observer.**

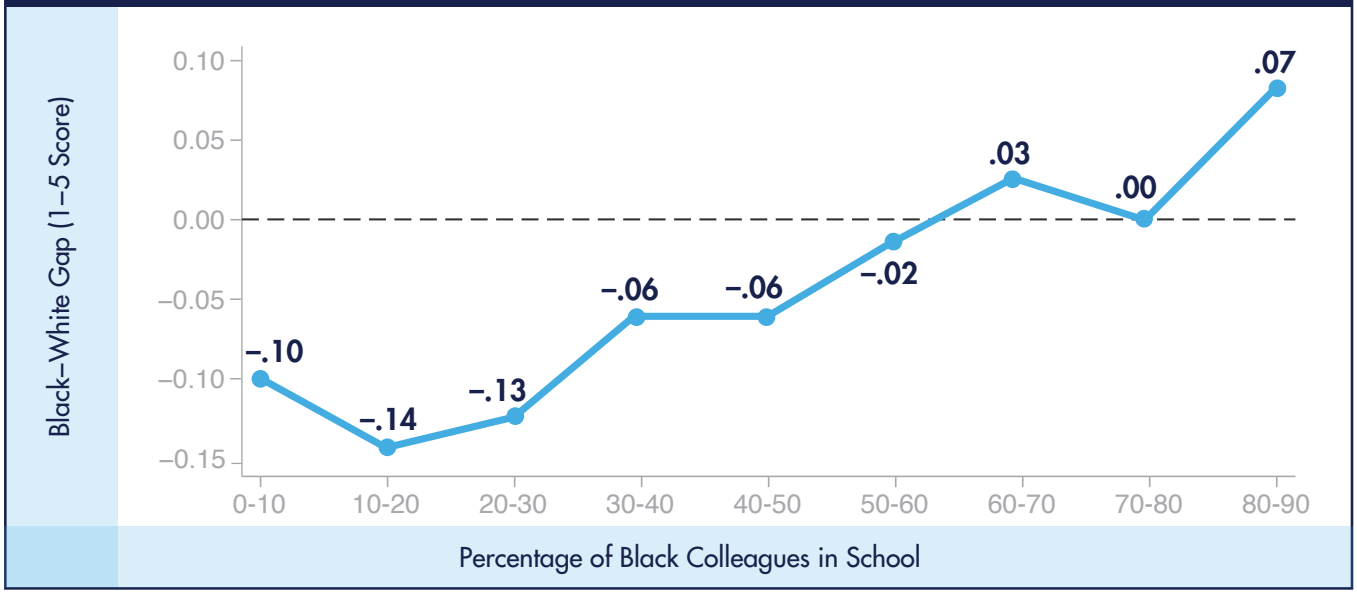
Black and White teachers work in different kinds of schools in Tennessee, and these differences may help explain the racial gaps we find in teachers' observation scores. For example, 71% of students in the typical Black teacher's school are eligible for free and reduced price lunches, a common measure of household poverty, compared to just 50% in the typical White teacher's school. If serving students in poverty affects teachers' observation scores (e.g., because meeting students' instructional needs is more challenging, because observation rubrics assign lower scores to instructional practices that are more effective with such students), Black teachers' tendency to be in schools with higher community poverty could lower their scores.

Comparing Black and White Teachers Across Different Types of Schools

To investigate the role of school context, we investigated how numerous school factors (e.g., student demographics, size, level, location) affected the estimated observation score gap between Black and White teachers. Among these, the racial composition of the school's teaching faculty emerged as especially important. We compared observation score gaps between Black and White teachers when Black teachers made up a very small percentage of the faculty to schools where they were more numerous, using regression to take other factors into account. As shown in Figure 4, score gaps are largest in schools where Black teachers are racially isolated—that is, where a Black teacher has few Black colleagues. **In schools with more racially diverse faculties, gaps are smaller, and, in fact, are erased (statistically) in schools where Black teachers make up about half a school's teachers.**

FIGURE 4

THE BLACK-WHITE GAP IN OBSERVATION SCORES IS LARGEST IN SCHOOLS WITH FEW BLACK TEACHERS AND DISAPPEARS IN SCHOOLS WHERE A LITTLE MORE THAN HALF OF THE TEACHERS ARE BLACK.



Comparing Black and White Teachers in the Same School in the Same Year

Further, given the importance of school context for how teachers are rated, we employed an additional statistical modeling strategy (called "fixed effects") to restrict comparisons to be only among Black and White teachers in the same school in the same school year. This restriction essentially sets school context to be the same among teachers in the analysis. **Even when comparing teachers in the same school in the same year, White teachers outscore Black teachers, on average, by approximately 0.08 points.**

Digging a little deeper, we also explored the degree to which within-school factors may explain the racial gaps in observation scores, such as the characteristics of students a teacher teaches and the race of a given teacher's observer. For example, similar to the differences between the types of schools in which Black and White teachers are most likely to work, even within the same school, Black and White teachers in Tennessee are not assigned the same types of students. Indeed, compared to their White colleagues, Black teachers are assigned students who are lower-achieving, more likely to have a past disciplinary history, have lower past attendance rates, and more likely to be Black, eligible for free and reduced-price lunch, and identified for special education services. Accounting for these student assignment differences reduces the estimated Black-White gap in observation scores by about 20% in schools with low fractions of Black teachers, where assignment gaps and observation score gaps are the largest.

Finally, we also considered differences in observation scores assigned by Black and White observers. The school leadership workforce in Tennessee is overwhelmingly White, so both Black and White teachers are more likely to have a White observer than a Black observer. Teachers often receive multiple classroom observations each year that are conducted by different raters (e.g., a principal and an assistant principal) who may have different racial identities, so we compared scores among Black and White teachers in the same school who were observed by these different raters and different times. This comparison uncovers an advantage to a teacher from being observed by an observer of the same race, though it is very small—about 0.02 points.

Taken together, we find that much—but not all—of the observation score gap between Black and White teachers can be explained by the differences in school context in which Black and White teachers work, and within-school factors, including the differences in the types of students that are assigned to Black and White teachers and whether they are observed by a rater of the same race. Although we can account for much of the race gap with what we can see in the data we have, there is still part of the gap that remains unexplained, warranting further investigation into drivers of differences in scores assigned to Black and White teachers.

Comparing Male and Female Teachers Across Different Types of Schools and Within the Same School

We conducted a similar set of analyses to understand the gap between male and female teachers. Unfortunately, this analysis yielded few clear insights. The gap between female and male teachers in the same school and year is virtually identical to the overall average gap, suggesting that school context does not play much of a role. Moreover, neither rater characteristics nor differential student assignment within schools explain the size of the female–male gap, and we do not find evidence that having a rater of the same gender influences a teacher's score. We do find some evidence that men and women tend to teach different grades and subjects—women are more prevalent in early grades, for example, and men are more likely to teach health and P.E.—and that accounting for these assignment differences shrinks the estimated gap by a few percentage points (because teachers teaching some content tend to receive higher ratings). However, beyond these potential differences by teaching assignment, we were not able to provide much by way of explanation for why female teachers' observation ratings are consistently higher than male teachers'.

CONCLUSION AND IMPLICATIONS

At a time of growing interest in increasing teacher diversity, it is valuable to understand factors that may affect the work experiences of teachers who are already underrepresented in the classroom and potentially push them to exit the teaching workforce. Our research finds that across evaluation systems and in schools with different characteristics across the state, Black and male teachers receive lower classroom observation scores than White and female teachers.

Importantly, Black-White and male-female score differences remain even when comparing similarly qualified teachers who perform the same according to other metrics, such as their value-added to student achievement. This finding raises concerns that the observation score gap reflects some form of systemic bias—that is, that Black and White (or male and female) teachers receive systematically different observation scores even when they have similar student achievement growth scores.

Bias in this sense does not require individual observers to be biased against particular groups of teachers. Nonrandom sorting of students within schools, which the research documents, could be a source of bias, if teachers tend to receive lower observation scores when they teach students who bring some challenges to the classroom—such as a history of disciplinary infractions, which may require a greater emphasis on classroom management—that other students do not bring. Another source of bias could be the observation rubrics themselves, which may give higher marks to teaching practices that some teachers are more likely to employ, even when other practices, employed by other teachers, are similarly effective.

Further investigation to understand the sources of these gaps is key to identifying solutions. Yet the presence of gaps across schools with so many different characteristics suggests the need for several next steps to address them to ensure that teachers across Tennessee are evaluated fairly.

Some Potential Next Steps for the State

- Continue exploring the sources of the Black-White gap in teacher observation scores.
- Ensure that the training observers receive to certify them to conduct classroom observations emphasize close application of the observation rubric.
- Audit approved observation rubrics to ensure that they reflect expectations for high-quality instructional practice for all teachers.

Some Potential Next Steps for School Districts

- Encourage school leaders to examine data on student placement in their schools each year to guard against systematic assignment of low-achieving students, students with a past history of disciplinary infractions, and so forth to Black teachers.
- Ensure that school leaders receive regular training on potential sources of bias in teacher evaluation.
- Regularly monitor observation scores assigned to White teachers and teachers of color to ensure that the observation process is being administered equitably.
- Consider observation scores as but one piece of evidence in making personnel decisions that affect teacher placement, retention, and compensation decisions.

END NOTES

1 The research reported in this brief was made possible (in part) by a grant from the Spencer Foundation (#202100068). The views expressed are those of the authors and do not necessarily reflect the views of the Spencer Foundation.

2 These mechanisms are discussed in a recent report, alongside a review of the evidence on teacher evaluation in Tennessee (Guthrie, Hernández, & Grissom, 2021).

3 This fraction is at least 50% for everyone but is higher for some teachers. For example, in 2020-21, for teachers in "untested" classrooms, observation ratings made up 70% of the final score.

4 This article is forthcoming in the peer-reviewed *Journal of Policy Analysis and Management.*

5 For details see https://www.tn.gov/content/tn/sbe/rules--policies-and-guidance/policies.html.

6 Approximately 98% of Tennessee teachers identify as either White or Black. Other groups make up very small fractions of teachers in the state. Given these small numbers and challenges in administrative data of accurately identifying some other teacher subgroups, the analysis summarized in this brief focuses only on Black and White teachers.

7 This measure is only available for teachers who teach subjects assessed by state tests. Tennessee's teacher evaluation system calculates a value-added score for such teachers called TVAAS. The researchers did not rely on this TVAAS score, however, instead calculating an alternative value-added score that addresses some statistical challenges with TVAAS. Details are available in Grissom and Bartanen (2021). In some models, the researchers also included a measure of teachers' value added to their students' attendance in the regression model to account for teachers' impacts on an important non-achievement outcome. In a small subset of districts who employed student surveys as part of their teacher evaluation system in some years, the models could also account for students' ratings of the teachers' instruction and classroom environment. The general patterns held with the inclusion of these additional performance measures.

8 Note that the magnitude of the Black-White gap within each evaluation system is larger than the Black-White gap that pools across all systems. While perhaps counterintuitive, this pattern is explained by the fact that Black teachers are more highly concentrated in districts using the TEM system, where average observation scores tend to be higher. Among teachers evaluated under TEM, 64% are Black, compared to 6%, 6.5%, and 2.5% for TEAM, COACH, and TIGER, respectively. The average observation score for TEM teachers (regardless of race) is 4.06, compared to 3.95 for teachers in other systems.

9 Teachers affect student outcomes other than achievement, and one possibility is that observation scores partly capture their performance in meeting other, non-achievement goals. To partially account for this possibility, the researchers estimated teachers' effects on an important non-achievement outcome, student attendance, and further compared teachers whose effects on attendance were the same. Even simultaneously setting teacher qualifications, effects on achievement, and effects on attendance to be the same, Black teachers receive observation scores that are significantly lower than White teachers, on average. Moreover, in a subsample of districts employing student surveys as part of the evaluation system, observation score gaps still remained among Black and White teachers whose practices were rated similarly by their students. See the full report (Grissom & Bartanen, 2021) for further details.

REFERENCES

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233–1267.

Drake, S., Auletto, A., & Cowen, J. M. (2019). Grading teachers: Race and gender differences in low evaluation ratings and teacher employment outcomes. *American Educational Research Journal, 56*(5), 1800–1833.

Grissom, J.A., & Bartanen, B. (2021). Potential race and gender biases in high-stakes teacher observations. In press, *Journal of Policy Analysis and Management.*

Guthrie, J.E., Hernández, M., & Grissom, J.A. (2021). *Teacher evaluation in Tennessee: What we have learned from a decade of research.* Nashville, TN: Tennessee Education Research Alliance. https://peabody.vanderbilt.edu/TERA/files/Teacher_Evaluation_Synthesis_FINAL.pdf

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, 43*(1), 60–82.