UWM UNIVERSITY of WISCONSIN MILWAUKEE

SREED Socially Responsible Evaluation in Education

Evidence of Discrimination and Bias in the Effectiveness Ratings Assigned to Wisconsin Educators of Color

WEERP Wisconsin Educator Effectiveness Research Partnership

June 2021

Curtis J. Jones, Leon Gilman, and Marlo Reeves, University of Wisconsin Milwaukee
Katharine Rainey, Wisconsin Department of Public Instruction (formerly)

This study is the first in a series examining the bias and discrimination affecting Wisconsin educators of color. In this study, we examine statewide effectiveness ratings data of over 55,000 educators for evidence of bias and discrimination. Bias can take many forms that diminish the ability of educators of color to succeed. They can be assigned more challenging classrooms with more underserved students or be expected to act as a disciplinarian for all students of color. They may also be viewed by their administrator less positively because of their race. This form of bias is often implicit or unconscious, rather than intentional. Discrimination occurs when an administrator acts on bias, regardless of its source. Acting on bias and assigning low effectiveness ratings to an educator of color is a form of discrimination.

Findings

The results of this study suggest ratings assigned to educators of color are discriminatory. Administrators view White female educators as the most effective, with Black and Asian male educators viewed as the least effective; 89% and 78% of White female educators are rated as more effective than the average Black and Latinx male educator, respectively. This was true even when comparing the ratings of educators with the same credentials and in the same schools. However, the performance appraisal process is likely just the tip of the iceberg regarding the negative impacts of bias on educators of color. The ratings reflect underlying biases that affect their experiences in ways that go deeper than the scope of this paper.

Evidence of Discrimination and Bias in the Effectiveness Ratings Assigned to Wisconsin Educators of Color

The Office of Socially Responsible Evaluation at the University of Wisconsin in Milwaukee has prepared this study as the first in a series exploring the systemic and interpersonal bias, and related discrimination, educators of color experience in Wisconsin schools, as reflected by the performance feedback provided to them as part of the Wisconsin Educator Effectiveness (EE) System. Within the EE System, administrators provide feedback to educators verbally, textually, and through effectiveness ratings. Administrators construct feedback that reflects their judgments about a teacher's effectiveness. However, there can be errors in what they perceive, how they apply the effectiveness rubric, and the feedback they provide. *Perceived* teacher effectiveness deviates from a teacher's true effectiveness when an evaluator assesses performance through the lens of his/her own biases. Interpersonal bias, such as sexism, racism, or other prejudices against a group, can negatively impact an evaluator's ability to accurately assess or recognize an educator's true performance. Interpersonal bias is not typically overt or intentional. Instead, it is driven by unconscious beliefs the individual is not even aware they have. Either way, a system that allows administrators to act on biased perceptions and rate educators of color as less effective is discriminatory.

The true effectiveness of an educator, absent interpersonal bias, reflects their actual effectiveness in conducting the assigned tasks as defined by an effectiveness rubric. However, the true measure of an educator's performance is not necessarily always fair. For instance, educators of color may be assigned to classrooms with more underserved students (Griffin & Tackie, 2016; Steinberg & Sartain, 2020) or be expected to fill other, unrelated, roles in their school, such as serving as the school's disciplinarian for students of color (Griffin & Tackie, 2016; Chapman, 2021). These conditions can diminish an educator's ability to succeed in their primary role as a classroom teacher. Thus, the evaluator may accurately apply the effectiveness rubric but miss critical contextual information that explains the educator's performance. The effectiveness ratings documented as part of the Wisconsin EE System provide a window into the magnitude of the interpersonal and systemic bias affecting educators of color.

In this report, we examine effectiveness ratings to measure the combined effects of systemic and interpersonal bias on perceptions of the effectiveness of Wisconsin educators of color. Differences in effectiveness ratings may represent a form of discrimination if they reflect biased perceptions about educators of color. A future study will examine the racial congruence between educators and evaluators to isolate the impacts of interpersonal and systemic bias on educators of color. Finally, to inform the efforts of Wisconsin schools for retaining educators of color and diversifying the workforce (Jones, 2019), a third report in this series will examine the potential for improving the retention of educators of color by reducing the systemic and interpersonal bias and discrimination affecting them.

Sources of bias affecting the performance of educators of color

Administrators generally perceive educators of color as less effective than their White colleagues (Campbell & Ronfeldt, 2018; Drake et al., 2019). A study of teacher performance in the Chicago Public Schools determined Black educators were more likely to be ranked in the lowest quartile of teacher performance when compared to White educators (Steinberg & Sartain, 2020). Systemic factors, such as administrators' expectations of educators of color to serve as a disciplinarian for students of color, or as the representative of their race or ethnicity (Chapman, 2021), may also negatively affect their effectiveness (Steinberg & Garrett, 2016). Further, as a group, educators of color serve more underserved schools and classrooms with more underserved students (Kalogrides et al., 2013), which affects their ability to demonstrate their skills (Campbell & Ronfeldt, 2018; Steinberg & Sartain, 2020).

Although Steinberg & Sartain (2020) suggested that interpersonal bias[1] played little role in the ratings assigned to Chicago educators of color, there is a broad body of research about performance reviews in business (Greenhaus & Parasuraman, 1993; Stauffer & Buckley, 2005; Constantine & Sue, 2007) and in education (Campbell & Ronfeldt, 2018; Drake et al., 2019; Jiang & Sporte, 2016) suggesting it does. In a qualitative study of 150 Black educators, educators reported administrators devalued them and viewed them as less educated and knowledgeable (Griffin & Tackie, 2016). In another study, Black women were more likely to be rated as less effective than White women, even when they were actually similarly effective in the classroom

[1] Of course, it is debatable if large-scale interpersonal bias represents interpersonal or systemic bias. In this paper, we treat interpersonal bias as separate from systemic bias.

(Campbell, 2020). There is evidence that interpersonal bias affects perceptions of male educators as well. The biased judgments regarding male teacher effectiveness may be due to evaluator perceptions that teaching is a more feminine profession, with males being less likely to fit into that stereotype (Wind et al., 2019). The influence of interpersonal bias is complicated and varies across a number of demographic aspects of educators. In a study in a large urban district, perceptions of teacher effectiveness varied by race, gender, age, and the intersection of all three. Evaluators were more likely to view Black educators, educators over the age of 50, and male educators as ineffective. (Bailey et al., 2016).

The current study examines racialized and gendered differences in the effectiveness ratings assigned to all Wisconsin educators participating in the Wisconsin EE System. Given the unique biases that male and female educators of color experience, we examine the intersection of race and gender in our analyses. The authors conceptualize any measured differences in effectiveness ratings as a form of discrimination resulting from both the interpersonal and systemic bias affecting educators of color.

The Teacher Evaluation System in Wisconsin – Educator Effectiveness

The Wisconsin EE system is based on research that teacher quality is the most important school factor for determining student achievement (McCaffrey et al., 2003; Rivkin et al., 2005). It is intended to promote the use of performance feedback to enhance the quality of teaching and student learning across the state. EE requires that schools provide ongoing, standards-based feedback to educators about their professional practices. Districts may use the Danielson Framework for Teaching (FfT) (Danielson, 2013), which is supported by the state, the Stronge Teacher Effectiveness Performance Standards (Stronge, 2002), which is supported by CESA 6 as the Effectiveness Project (EP), or apply to use another equivalent rubric. Approximately two-thirds of Wisconsin educators are provided feedback according to the FfT and one-third according to Stronge Standards used in the EP.

Educator Effectiveness Rubrics

The Danielson Framework for Teaching (FfT) measures educator effectiveness across 22 components and four domains (Figure 1). These domains include Planning & Preparation, Classroom Environment, Instruction, and Professional Responsibilities. An administrator

typically rates educators on a one-to-four scale from Unsatisfactory (1), Basic (2), Proficient (3), or Distinguished (4) on each of the 22 components that comprise these domains.

Schools using the Stronge Teacher Effectiveness Performance Standards, as part of the EP system, provide educators performance feedback specific to six standards. These include Professional Knowledge, Instructional Planning, Instructional Delivery, Assessment for/of Learning, Learning Environment, and Professionalism (Figure 2). Educators evaluated as part of the EP typically receive ratings on a one to four scale from Unacceptable (1), Developing/Needs Improvement (2), Effective (3), and Distinguished (4).

For the purposes of this study, we created an overall FfT and Stronge (EP) rating. To calculate the overall FfT rating, we averaged each of the 22 components into four domain scores, which we then averaged into an overall FfT rating. For the EP model, we averaged the six Stronge Standards into an overall rating. Correlations between FfT Domains ratings and the overall FfT ratings were near .90 (Table 1; Appendix A). Likewise, the correlations between the six Stronge Standards and the overall rating were 0.70 (Table 2; Appendix A). Thus, any discussion of ratings by overall scores should be representative of domains and component ratings. However, our use of overall ratings is for research purposes only and does not reflect the typical practice of how educators receive performance feedback.

Figure 1: The 22 components and four domains that comprise the Danielson Framework for Teaching (FfT) rubric.

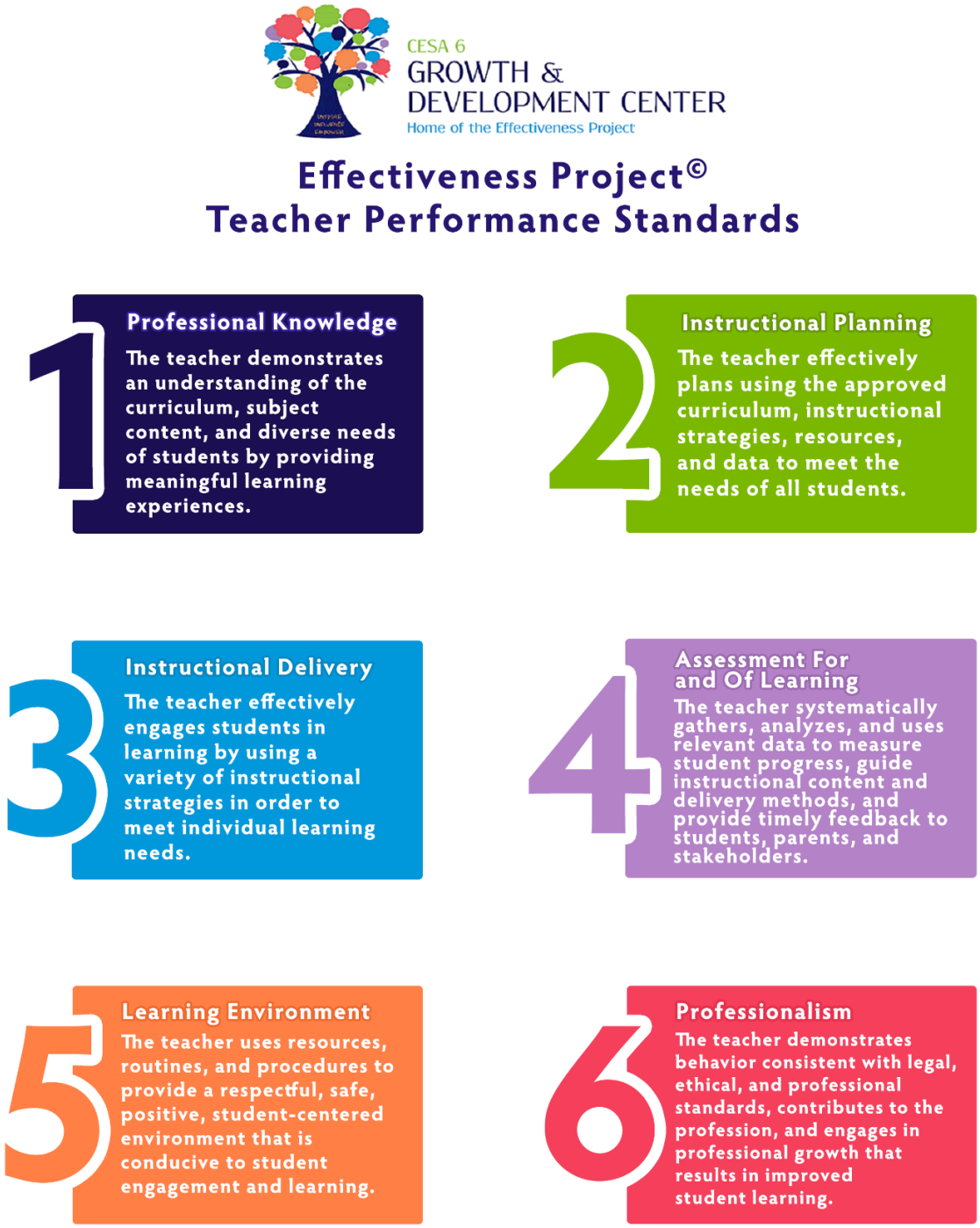

Figure 2: The six Stronge Teacher Effectiveness Performance Standards used in the CESA 6 Effectiveness Project (EP).

Current Study

This study examines teacher and school factors associated with the effectiveness ratings assigned to male and female educators from different racial/ethnic groups. The purpose of this study is to uncover evidence of discrimination by measuring the combined impact of interpersonal and systemic bias on the effectiveness ratings assigned to educators of color. Subsequent studies will work to disentangle the sources of any measured bias and discrimination this study uncovers and to understand its impact on the retention of educators of color.

This study includes all Wisconsin classroom educators who received effectiveness ratings at least once between the first year of statewide implementation (2014-15 school year) and the 2019-20 school year. We matched effectiveness ratings with state educator records that include district and school assignment, years of experience in the district, total years of experience in public education, educational attainment, race, and gender.

The Frontline Education (FE) data system is the primary tool used to document local EE processes. During the 2014-15 and 2015-16 school years, only EP schools used FE. During these two years, districts using the FfT rubric documented effectiveness ratings in another data system called Teachscape. Of districts using the FfT, only large urban school districts documented their ratings during the transition from Teachscape to FE in 2015-16.

Recorded FfT data includes performance ratings across each of its 22 components. Recorded EP data includes performance ratings across each of the six Stronge standards. We successfully matched 93,299 (94.5%) of the 98,685 FfT or EP ratings to state educator demographic information.[2] The 93,299 matched ratings represent 55,963 educators. 51,945 (55.7%) matched ratings used the FfT, and 41,354 (44.3%) used the Stronge Standards (EP) as their rubric. The matched FfT ratings represented educators in 1,384 schools, and Stronge Standards (EP) represented educators in 840 schools.

Publicly available ratings data

The authors developed two publicly available dashboards to accompany this report, one that allows for examining Danielson FfT ratings assigned to educators

[2] Detailed educator demographic information are included in Table 3, 4, and 5 (Appendix B).

(https://uwm.edu/sreed/educator-effectiveness/danielson-fft-ratings-dashboard/), and the other for examining Stronge Performance Standards assigned to educators in the EP (https://uwm.edu/sreed/educator-effectiveness/cesa-6-effective-project-ep-historical-stronge-rubric-ratings/). These dashboards enable the user to explore effectiveness ratings by teacher demographic and background characteristics, school types, and community characteristics. The current study focuses mostly on the relationships of race and gender with effectiveness ratings. We understand the potential for additional powerful and interesting analyses of effectiveness ratings. It is our hope that these dashboards will empower researchers, educators, and policymakers to explore the data themselves.

Results

What ratings did evaluators assign to male and female educators from different racial/ethnic backgrounds?

Across all educators, the average FfT rating was 3.08 with a standard deviation of 0.33. White educators received the highest ratings (3.10) and Black educators (2.81) the lowest (Table 6). The largest gap between any group was 0.41 scale points between the ratings assigned to Black male (2.71) and White female educators (3.12). **Considering the FfT standard deviation of 0.33, this difference of 0.41 scale points reflects a difference of 1.24 standard deviations, which suggests 89% of White female educator FfT ratings were higher than the average rating for Black male educators.** Other comparisons between the FfT effectiveness ratings of other groups and White educators suggest:

- 79% of White female educators were rated as more effective than the average Black female educator.
- 78% of White female educators were rated as more effective than the average Asian male educator.
- 77% of White female educators were rated as more effective than the average Latinx male educator.

Regarding Stronge standards (EP), the average rating was 3.19 with a standard deviation of 0.35. White educators again received the highest ratings (3.20) with Black (3.07) and Asian educators (3.06) rated lowest (Table 6). Among gender subgroups, White female educators received the

highest ratings (3.21) and Asian male educators the lowest (2.94). **This gap of 0.27 scale points reflects a 0.78 standard deviation difference, suggesting 78% of White female educators received higher EP ratings than the average Asian male educator. Similarly, 71% of White female educators were rated as more effective than the average Black male educator.**

Table 6: Overall ratings assigned to educators with different racial and gender identities.

	FfT			Stronge (EP)		
	Average	SD	Ratings	Average	SD	Ratings
Female educators						
American Indian or Alaska Native	3.07	0.31	126	3.12	0.41	54
Asian	3.01	0.37	429	3.11	0.41	112
Black	2.85	0.45	1,187	3.10	0.36	58
Latinx	3.03	0.34	1,206	3.11	0.32	219
Native Hawaiian or other Pacific Islander	3.07	0.22	24	3.07	0.26	10
Two or more races	2.97	0.35	148	3.19	0.35	67
White	3.12	0.31	35,690	3.21	0.35	29,797
Male educators						
American Indian or Alaska Native	2.94	0.38	37	3.07	0.25	30
Asian	2.87	0.40	145	2.94	0.34	47
Black	2.71	0.44	464	3.02	0.38	39
Latinx	2.88	0.40	449	3.12	0.41	73
Native Hawaiian or other Pacific Islander	3.04	0.21	6	3.37	0.22	3
Two or more races	2.91	0.43	65	3.11	0.30	33
White	3.04	0.34	11,969	3.15	0.34	10,812
Overall						
American Indian or Alaska Native	3.04	0.33	163	3.11	0.36	84
Asian	2.98	0.38	574	3.06	0.40	159
Black	2.81	0.45	1,651	3.07	0.37	97
Latinx	2.99	0.36	1,655	3.11	0.35	292
Native Hawaiian or other Pacific Islander	3.06	0.21	30	3.14	0.28	13
Two or more races	2.95	0.37	213	3.16	0.34	100
White	3.10	0.32	47,659	3.20	0.35	40,609

What are the differences in ratings assigned to male and female educators of color compared to the ratings assigned to White female educators within the same school with the same credentials?

While the unadjusted results suggest race is strongly related to perceptions of teachers' effectiveness, it is unclear to what extent these differences reflect differences between "like" educators in similar educational settings. We used statistical modeling (see Appendix C) to make a more precise comparison between the racial and gender differences in FfT and Stronge (EP) ratings assigned to educators. These analyses included a population of 50,605 Fft and 41,164 Stronge (EP) ratings. In these models, we account for the school within which educators work and their experience and education. We know from Appendix B that educators of color work in different schools than White educators. The inclusion of fixed school effects in our analyses controls for any observable and unobservable differences between schools that might bias ratings comparisons between racial and gender groups. The results of these analyses answer the question, ***"What are the differences in ratings assigned to male and female educators of color compared to the ratings assigned to White female educators <u>with the same credentials in the same schools</u>?"***

Looking within schools at educators matched by experience and education, effectiveness ratings differences between White female educators and other educator groups are reduced but still sizable (Table 7; Appendix D). Consistent with statewide ratings presented in Table 6, the largest group difference was still between White female and Black male educators. Adjusting ratings by the school and teacher credentials reduced the magnitude of the difference between Black male and White female educators by about half, from 0.41 to 0.20. This suggests that <u>within the same school</u>, you would expect a Black male teacher to be rated as 0.20 points lower than a White female with the same experience and education. **This difference of 0.20 scale points reflects 0.61 standard deviations and suggests evaluators rated 73% of White female educators higher than Black male educators in the same school with the same credentials. Similarly, evaluators rated 68% of White females as more effective than similarly credentialed Asian male educators.**

Consistent with the unadjusted results presented in Table 6, after adjusting for school and teacher credentials, a number of groups still received lower Stronge (EP) ratings than White female

educators (Table 8; Appendix D). The largest gaps were observed between the ratings assigned to Black (0.14) and Asian (0.25) male educators. Contrary to FfT results, school and teacher adjustments did little to reduce the size of the rating gaps between White female educators and these groups. The (0.14) point difference between White female and Black male educators reflects a 0.4 standard deviation difference. This difference suggests that 66% of White female educators received higher ratings than male educators with the same credentials in the same school. **The 0.25 point difference between White female and Asian male educators reflects 0.71 standard deviations. This difference suggests that 76% of White female educators received higher ratings than the typical Asian male educator in their same school with the same credentials.**

We conducted a robustness check to account for some of the limitations of using fixed school effects. Many educators of color are the only educator of color in their school. Because of this, school fixed effects may over-identify the model and result in many educators of color not factoring into estimates of relationships between race and effectiveness ratings. For instance, there were 280 schools with at least one Black educator rated with the FfT. In 76 of these schools, only one rating of a Black educator occurred. As an alternative to including fixed school effects, these robustness check models account for the characteristics of schools, i.e., their racial, economic, and demographic composition, the type of community the school is within, and the type of school (elementary, middle, or high). Including school characteristics instead of school fixed effects in our analysis allows for a more inclusive analysis. These models answer the question, ***"What are the differences in ratings assigned to male and female educators of color compared to the ratings assigned to White female educators with the same credentials in similar schools?"*** The results of these analyses are consistent with the fixed school effects models and again suggest Asian male and Black male educators were viewed as the least effective according to both the FfT and Stronge Standards (EP) (Tables 9 and 10; Appendix E).

Discussion

This study provides evidence that effectiveness ratings are discriminatory, reflecting biases affecting Wisconsin educators of color. Educators of color, as a group, are viewed by school administrators as less effective than White educators. An evaluator would likely perceive a White educator as more effective than a Black educator in the same school with the same education and experience. This was especially apparent when comparing perceptions of White female educators with Black and Asian male educators. These differences in perceptions are due to a number of structural or systemic factors related to different histories and circumstances affecting students in classrooms and the roles educators may be asked to fill in their school that diminish their ability to succeed. Interpersonal factors, such as implicit biases and administrator expectations regarding what an effective educator's background should be, also contribute.

These findings do not suggest removing effectiveness ratings would fix the problem. The ratings assigned to educators are a form of discrimination, resulting from the biased perceptions of administrators. The problem lies in the circumstances impacting educators of color and the biased perceptions of administrators. Ratings provide an opportunity to measure one consequence of that bias. Remove the ratings from EE, and the bias remains, affecting the experiences of educators of color in deeper ways that go beyond the scope of this study. However, removing the ratings from the EE process would make it easier to ignore the systemic and interpersonal biases and discrimination affecting Wisconsin educators of color.

The results of this study provide an omnibus measure of bias affecting male and female educators of color. However, they do not provide for the ability to disentangle the impact of systemic bias from interpersonal bias. We will work to isolate the effect of interpersonal bias in our next study by comparing the ratings assigned to educators of color when their evaluator is and is not from their same racial background. After accounting for this, what remains may reflect the effects of systemic bias related to such issues as the roles educators of color are asked to fill in their school (Chapman, 2021) and the histories of students who are assigned to educators (Kalogrides et al., 2013). As part of this study, we will also compare the textual performance feedback provided to educators of color and White educators, depending on the race of their evaluator. Our final planned study in this line of research will identify schools that have demonstrated less bias and discrimination in the performance appraisal process of educators of

color. We will use these schools to measure the impact of bias and discrimination on educator retention and learn from them possible strategies for reducing them. Through this work, we hope to educate schools on how to organize to be more inclusive and positive settings that better support, empower, and retain educators of color.

Appendix A – Correlations of Stronge Standards and FfT Domain ratings

Table 1: FfT domains and overall rating correlations.

	1.	2.	3.	4.	5.
1. Planning & Preparation	-				
2. Classroom Environment	.695	-			
3. Instruction	.781	.746	-		
4. Professional Responsibilities	.764	.654	.701	-	
5. Overall FfT Rating*	.906	.878	.905	.872	-

* Calculated solely for the purpose of this study.

Table 2: Overall Stronge (EP) correlations with the six standards ratings.

	1.	2.	3.	4.	5.	6.	7.
1. Professional Knowledge	-						
2. Instructional Planning	.443	-					
3. Instructional Delivery	.464	.448	-				
4. Assessment For/Of Learning	.407	.459	.393	-			
5. Learning Environment	.404	.398	.509	.344	-		
6. Professionalism	.471	.392	.386	.346	.390	-	
7. Overall EP Rating*	.741	.726	.748	.671	.723	.702	-

* Calculated solely for the purpose of this study.

Appendix B – Factors associated with ratings included in the study

Factors accounted for in this study included teacher, school, and community characteristics. Including these factors help isolate and contextualize any measured gendered and racialized differences in effectiveness ratings.

Teacher characteristics included in our analyses:

- Total teacher experience categorized into three groups (10.5 or more years, 3.5 to 10.0 years, and 0.5 to 3.0 years),
- Highest educational degree received dichotomized into two groups (having a master's degree or a bachelor's degree),
- Race/ethnicity (American Indian or Alaska Native, Asian, Black or African American, Hispanic /Latino, Native Hawaiian or Other Pacific Islander, two or more races, or White), and
- Gender (Male or female, as is defined by DPI data records. DPI records do not provide educators the opportunity to identify themselves in other ways.).

Table 3 presents the number of effectiveness ratings assigned to educators from different racial and ethnic groups broken down by other teacher characteristics. Consistent with our previous research that demonstrates teachers of color leave the profession at faster and greater rates than their White colleagues due to issues of bias (Jones, 2019), White educators are more likely to be experienced educators and slightly more likely to have a master's degree. Considering the connection between experience, education, and effectiveness (Goldhaber & Brewer, 1997; Kini & Podolsky, 2016), our statistical comparisons of effectiveness ratings assigned to different racial groups account for their education and experience. Of course, accounting for experience and education adjusts for some of how bias impacts teachers of color. As such, we present both effectiveness ratings adjusted by and not adjusted by experience and education.

School and community types included in our analyses:

- Community type (Rural, Suburb, Town, and Urban) and
- School type (elementary, middle, and high school).

School and community types are displayed in Table 4. White and American Indian or Alaska Native educators worked across all four types of communities. Black, Latinx, and Asian educators worked mostly in urban communities.

School characteristics included in our analyses:

- School size (students enrolled),
- The percentage of students who have an IEP,
- The percentage of students who are economically disadvantaged,
- The percentage of students who are English Learners (EL),
- The percentage of students who are Black,
- The percentage of students who are Latinx, and
- The percentage of students who are White.

Table 5 presents the characteristics of schools where educators received their ratings. Educators of color tended to work in schools with a higher percentage of students identified as economically disadvantaged. Asian, Hispanic/Latino, Native Hawaiian or Other Pacific Islander educators worked in schools with more English Learners. Educators tended to work in schools with more students from their same racial/ethnic background.

Table 3: Frequency of ratings assigned to different groups of educators by race/ethnicity of educator.

	FfT							EP						
	Years of Experience			Degree		Gender		Years of Experience			Degree		Gender	
	0.5 - 3.0	3.5 - 10.0	10.5+	Bachelor	Master	Female	Male	0.5 - 3.0	3.5 - 10.0	10.5+	Bachelor	Master	Female	Male
American Indian or Alaska Native	44	50	69	86	66	126	37	26	18	39	52	25	54	30
	27.0%	30.7%	42.3%	56.6%	43.4%	77.3%	22.7%	31.3%	21.7%	47.0%	67.5%	32.5%	64.3%	35.7%
Asian	173	161	238	309	235	429	145	55	45	59	100	59	112	47
	30.2%	28.1%	41.6%	56.8%	43.2%	74.7%	25.3%	34.6%	28.3%	37.1%	62.9%	37.1%	70.4%	29.6%
Black	641	349	657	894	560	1,187	464	38	31	28	65	31	58	39
	38.9%	21.2%	39.9%	61.5%	38.5%	71.9%	28.1%	39.2%	32.0%	28.9%	67.7%	32.3%	59.8%	40.2%
Latinx	583	508	564	1,001	539	1,206	449	113	83	95	193	94	219	73
	35.2%	30.7%	34.1%	65.0%	35.0%	72.9%	27.1%	38.8%	28.5%	32.6%	67.2%	32.8%	75.0%	25.0%
Native Hawaiian or other Pacific Islander	11	11	8	18	12	24	6	5		5	8	5	10	
	36.7%	36.7%	26.7%	60.0%	40.0%	80.0%	20.0%	38.5%		38.5%	61.5%	38.5%	76.9%	
Two or more races	79	74	60	148	61	148	65	37	28	37	66	34	68	34
	37.1%	34.7%	28.2%	70.8%	29.2%	69.5%	30.5%	36.3%	27.5%	36.3%	66.0%	34.0%	66.7%	33.3%
White	11,527	12,498	23,610	25,822	20,854	35,690	11,969	9,440	10,569	20,724	23,165	17,447	29,935	10,854
	24.2%	26.2%	49.6%	55.3%	44.7%	74.9%	25.1%	23.2%	25.9%	50.9%	57.0%	43.0%	73.4%	26.6%

Table 4: Frequency of ratings assigned to different school and community types by race/ethnicity of educator.

	FfT							EP						
	Community Type				School Type			Community Type				School Type		
	Town	Suburb	Urban	Rural	ES	MS	HS	Rural	Suburb	Town	Urban	ES	MS	HS
American Indian or Alaska Native	21	23	73	45	100	22	35	23	29	6	26	37	12	35
	13.0%	14.2%	45.1%	27.8%	63.7%	14.0%	22.3%	27.4%	34.5%	7.1%	31.0%	44.0%	14.3%	41.7%
Asian	20	67	460	24	317	73	152	37	67	38	15	69	31	53
	3.5%	11.7%	80.6%	4.2%	58.5%	13.5%	28.0%	23.6%	42.7%	24.2%	9.6%	45.1%	20.3%	34.6%
Black	5	84	1,512	21	946	134	378	12	48	33		34	20	35
	0.3%	5.2%	93.2%	1.3%	64.9%	9.2%	25.9%	12.4%	49.5%	34.0%		38.2%	22.5%	39.3%
Latinx	38	183	1,363	60	1016	197	363	106	72	42	66	95	62	120
	2.3%	11.1%	82.9%	3.6%	64.5%	12.5%	23.0%	37.1%	25.2%	14.7%	23.1%	34.3%	22.4%	43.3%
Native Hawaiian or other Pacific Islander		5	20		15	6	6					9		
		17.9%	71.4%		55.6%	22.2%	22.2%					69.2%		
Two or more races	11	42	130	23	91	33	65	28	26	18	29	44	24	34
	5.3%	20.4%	63.1%	11.2%	48.1%	17.5%	34.4%	27.7%	25.7%	17.8%	28.7%	43.1%	23.5%	33.3%
White	6,735	11,052	21,362	8,221	24,352	8,086	13,194	12,005	9,929	5,326	13,150	19,131	8,044	12,555
	14.2%	23.3%	45.1%	17.4%	53.4%	17.7%	28.9%	29.7%	24.6%	13.2%	32.5%	48.2%	20.2%	31.6%

Table 5: Average Student body characteristics of schools where rated educators worked by race/ethnicity of educator.

	FfT								EP							
	Students	SwD	ED	EL	Black	Latinx	White	Ratings	Students	SwD	ED	EL	Black	Latinx	White	Ratings
Amer Indian or Alaska Native	552.1	16.4%	56.8%	7.7%	13.2%	15.0%	46.8%	163	597.3	15.1%	40.7%	2.5%	3.2%	9.0%	66.7%	84
Asian	687.5	16.3%	58.0%	12.9%	25.1%	17.2%	42.7%	573	839.8	13.4%	38.9%	5.7%	5.1%	11.9%	74.4%	159
Black	612.6	21.3%	79.1%	7.0%	64.4%	14.4%	14.1%	1,646	750.0	13.1%	36.2%	3.7%	10.4%	11.9%	69.3%	97
Latinx	712.1	16.8%	67.5%	23.8%	19.3%	43.0%	29.5%	1,653	657.4	13.2%	42.6%	8.6%	3.7%	19.2%	71.0%	291
Native Hawaiian or other Pacific Islander	648.6	15.2%	48.5%	11.8%	18.9%	16.2%	53.0%	30	514.6	13.8%	38.6%	12.9%	2.1%	17.4%	73.2%	13
Two or more races	697.2	16.4%	55.9%	9.9%	24.4%	18.9%	45.9%	213	665.0	14.2%	39.2%	3.9%	3.9%	10.1%	74.4%	102
White	644.7	14.5%	44.8%	7.8%	12.0%	13.7%	63.9%	47,553	583.6	13.6%	35.5%	3.6%	2.7%	8.8%	82.0%	40,726

Appendix C – Modeling approach

Two statistical modeling approaches were used to ***adjust*** effectiveness ratings and isolate differences in ratings assigned to educators from different racial and ethnic backgrounds. In both, Generalized Linear Modeling (GLM) with robust standard errors were used to predict teacher performance ratings. Our first model used school fixed effects and was expressed as:

$$Y_i = \beta_0 + \sum_{m=1}^{M} \beta_{1.m} X_{mi} + \beta_2 (Race \times Gender_i) + \sum_{m=1}^{M-1} \beta_{3.m}\, Year_{mi} + \sum_{m=1}^{M-1} \beta_{4.m}\, School_{mi} + \epsilon_i \quad (1)$$

where Y_i is the outcome (overall FfT or Overall Stronge (EP) rating) for the ith person, β_0 is the intercept; β_1 is the effects of teacher characteristics; X_{mi} is the mth of M additional covariates representing teacher characteristics (experience and highest degree earned); β_2 is the impact of the race/ethnicity by gender interaction effect for the ith person, β_3 is the effect of year the rating was assigned, β_4 is the effect of school, and ϵ_i is error term for the ith person.

Another set of statistical models were used that included school, community and student body characteristics in the model instead of school fixed effects. The purpose of this was to include all educators in our analysis, even if they were the only educator of color in a school. This model was expressed as,

$$Y_i = \beta_0 + \sum_{m=1}^{M} \beta_{1.m} X_{mi} + \beta_2 (Race \times Gender_i) + \sum_{m=1}^{M-1} \beta_{3.m}\, Year_{mi} + \sum_{m=1}^{M-1} \beta_{4.m}\, X_{mi} + \epsilon_i \quad (2)$$

where β_4 is the effect of school, community, and student body characteristics and X_{mi} is m^{th} of M additional covariates representing these characteristics (school type, community type, enrollment, percentage of students that have a disability, are economically disadvantaged, are English Learners, are Black, are Latinx, or are White).

Appendix D – Model Results Accounting for Fixed School Effects

Table 7: Model results of overall FfT rating for different racial and gender demographic groups with the same credentials in the same school.

Group	Rating Diff	Std. Error	Wald Chi-Sq.	*df*	*p*
Two or more races male educators	**-0.139**	**0.044**	**10.155**	**1**	**0.001**
Two or more races female educators	**-0.059**	**0.024**	**6.304**	**1**	**0.012**
Native Hawaiian or Other Pacific Islander male educators	-0.039	0.114	0.117	1	0.732
Native Hawaiian or Other Pacific Islander female educators	0.030	0.036	0.700	1	0.403
Latinx male educators	**-0.107**	**0.017**	**38.327**	**1**	**0.000**
Latinx female educators	-0.006	0.010	0.356	1	0.550
Black male educators	**-0.205**	**0.020**	**101.360**	**1**	**0.000**
Black female educators	**-0.090**	**0.014**	**42.804**	**1**	**0.000**
Asian male educators	**-0.155**	**0.031**	**25.446**	**1**	**0.000**
Asian female educators	**-0.050**	**0.017**	**8.494**	**1**	**0.004**
American Indian or Alaska Native male educators	-0.093	0.057	2.690	1	0.101
American Indian or Alaska Native female educators	-0.019	0.024	0.664	1	0.415
White male educators	**-0.067**	**0.003**	**427.588**	**1**	**0.000**

Notes: The reference group is White female educators.

Bold = $p < 0.05$.

Table 8: Model results of overall Stronge (EP) rating for different racial and gender demographic groups with the same credentials in the same school.

Group	Rating Diff	Std. Error	Wald Chi-Sq.	*df*	*p-value*
Two or more races male educators	-0.046	0.051	0.802	1	0.371
Two or more races female educators	0.004	0.035	0.017	1	0.898
Native Hawaiian or Other Pacific Islander male educators	0.191	0.123	2.410	1	0.121
Native Hawaiian or Other Pacific Islander female educators	-0.070	0.077	0.825	1	0.364
Latinx male educators	-0.039	0.041	0.914	1	0.399
Latinx female educators	**-0.050**	**0.021**	**5.714**	**1**	**0.017**
Black male educators	**-0.135**	**0.056**	**5.903**	**1**	**0.015**
Black female educators	-0.044	0.041	1.109	1	0.292
Asian male educators	**-0.252**	**0.050**	**25.892**	**1**	**0.000**
Asian female educators	**-0.089**	**0.037**	**5.706**	**1**	**0.017**
American Indian or Alaska Native male educators	**-0.085**	**0.039**	**4.760**	**1**	**0.029**
American Indian or Alaska Native female educators	-0.031	0.045	0.459	1	0.498
White male educators	**-0.064**	**0.004**	**315.699**	**1**	**0.000**

Notes: The reference group is White female educators.

Bold = $p < 0.05$.

Appendix E – Model Results Accounting for School Characteristics

Table 9: Model results of overall FfT rating for different racial and gender demographic groups with the same credentials in similar schools.

Group	Rating Diff	Std. Error	Wald Chi-Sq.	*df*	*p-value*
Two or more races male educators	**-0.151**	**0.051**	**8.604**	**1**	**0.003**
Two or more races female educators	**-0.055**	**0.026**	**4.501**	**1**	**0.034**
Native Hawaiian or Other Pacific Islander male educators	-0.073	0.095	0.595	1	0.440
Native Hawaiian or Other Pacific Islander female educators	-0.007	0.044	0.024	1	0.877
Latinx male educators	**-0.126**	**0.018**	**47.105**	**1**	**0.000**
Latinx female educators	**-0.024**	**0.010**	**5.519**	**1**	**0.019**
Black male educators	**-0.233**	**0.022**	**108.083**	**1**	**0.000**
Black female educators	**-0.102**	**0.015**	**44.868**	**1**	**0.000**
Asian male educators	**-0.162**	**0.033**	**24.785**	**1**	**0.000**
Asian female educators	**-0.054**	**0.018**	**9.319**	**1**	**0.002**
American Indian or Alaska Native male educators	-0.081	0.056	2.143	1	0.143
American Indian or Alaska Native female educators	-0.029	0.026	1.241	1	0.265
White male educators	**-0.063**	**0.004**	**322.686**	**1**	**0.000**

Notes: The reference group is White female educators.

Bold = $p < 0.05$.

Table 10: Model results of overall Strong (EP) rating for different racial and gender demographic groups with the same credentials in similar schools.

Group	Rating Diff	Std. Error	Wald Chi-Sq.	*df*	*p-value*
Two or more races male educators	-0.064	0.046	1.942	1	0.163
Two or more races female educators	0.015	0.038	0.158	1	0.691
Native Hawaiian or Other Pacific Islander male educators	0.027	0.117	0.054	1	0.816
Native Hawaiian or Other Pacific Islander female educators	-0.089	0.074	1.429	1	0.232
Latinx male educators	-0.036	0.044	0.688	1	0.407
Latinx female educators	**-0.038**	**0.021**	**3.463**	**1**	**0.063**
Black male educators	**-0.186**	**0.058**	**10.413**	**1**	**0.001**
Black female educators	**-0.100**	**0.047**	**4.590**	**1**	**0.032**
Asian male educators	**-0.246**	**0.050**	**24.382**	**1**	**0.000**
Asian female educators	**-0.089**	**0.038**	**5.474**	**1**	**0.019**
American Indian or Alaska Native male educators	**-0.101**	**0.044**	**5.275**	**1**	**0.022**
American Indian or Alaska Native female educators	-0.035	0.055	0.406	1	0.524
White male educators	**-0.063**	**0.004**	**263.234**	**1**	**0.000**

Notes: The reference group is White female educators.

Bold = $p < 0.05$.

References

Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). *Teacher Demographics and Evaluation: A Descriptive Study in a Large Urban District*.

Campbell, S. L. (2020). Ratings in black and white: a quantcrit examination of race and gender in teacher evaluation reform. *Race Ethnicity and Education*, 1-19. DOI: 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*, *55*(6), 1233–1267.

Chapman, A. (2021). *Opening Doors: Strategies for Advancing Racial Diversity in Wisconsin's Teacher Workforce*. Wisconsin Policy Forum. https://wispolicyforum.org/wp-content/uploads/2021/03/OpeningDoors_FullReport.pdf

Constantine, M. G., & Sue, D. W. (2007). Perceptions of racial microaggressions among black supervisees in cross-racial dyads. Journal of Counseling Psychology, 54(2), 142–153.

Danielson, C. (2013). *The Framework for Teaching Evaluation Instrument, 2013 Instructionally Focused Edition*. The Danielson Group.

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.

Goldhaber, D. D., & Brewer, D. J. (1996). *Evaluating the effect of teacher degree level on educational performance.*

Greenhaus, J. H., & Parasuraman, S. (1993). Job performance attributions and career advancement prospects: An examination of gender and race effects. Organizational Behavior and Human Decision Processes, 55(2), 273-297.

Griffin, A., & Tackie, H. (2016). Through Our Eyes: Perspectives and Reflections from Black Teachers. Washington, DC: The Education Trust.

Jiang, J. Y., & Sporte, S. E. (2016). Teacher Evaluation in Chicago: Differences in Observation and Value-Added Scores by Teacher, Student, and School Characteristics. Research Report. *University of Chicago Consortium on School Research*.

Jones, C. J. (2019). *Race, Relational Trust, and Teacher Retention*. Wisconsin Educator Effectiveness Research Partnership (WEERP). Retrieved online at: https://uwm.edu/sreed/wp-content/uploads/sites/502/2019/11/WEERP-Brief-Nov-2019-Race-Relational-Trust-and-Teacher-Retention.pdf

Kini, T., & Podolsky, A. (2016). Does Teaching Experience Increase Teacher Effectiveness? A Review of the Research. Palo Alto: Learning Policy Institute. This report can be found at https://learningpolicyinstitute.org/our-work/publications-resources/does-teaching-experience-increase-teacher-effectiveness-review-research.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability. [Monograph.]*. RAND Corporation.

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535–566.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Stauffer, J. M., & Buckley, M. R. (2005). The Existence and Nature of Racial Bias in Supervisory Ratings. Journal of Applied Psychology, 90(3), 586–591.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317.

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. https://doi.org/10.3102/0162373720970204

Stronge, J. H. (2002). *Qualities of Effective Teachers*. Association for Supervision and Curriculum Development.

Wind, S. A., Jones, E., Bergin, C., & Jensen, K. (2019). Exploring patterns of principal judgments in teacher evaluation related to reported gender and years of experience. *Studies in Educational Evaluation*, *61*, 150–158.

This project is part of the Wisconsin Educator Effectiveness partnership between the Wisconsin Department of Public Instruction, the University of Wisconsin in Milwaukee, and the University of Wisconsin in Madison.

We would like to the following individuals for their support and feedback on this report:

Sarah Archibald, University of Wisconsin Madison
Sheila Briggs, Wisconsin Department of Public Instruction
William Cannon, Wisconsin Department of Public Instruction
Cathy Clarksen, CESA 6
Annalee Good, University of Wisconsin Madison
Kim Hill, CESA 10
Jacob Hollnagel, Wisconsin Department of Public Instruction
Steve Kimball, University of Wisconsin Madison
Emily Kite, University of Wisconsin Madison
Scott Prinster, Wisconsin Department of Public Instruction

Curtis J. Jones is the Director of the Office of Socially Responsible Evaluation in Education (SREED) at the University of Wisconsin Milwaukee (UWM) and Co-Director of the Wisconsin Educator Effectiveness Research Partnership (WEERP).

Leon Gilman is an Evaluation/Research Associate in SREED at UWM.

Marlo Reeves is a Senior Evaluation/Research Associate in SREED at UWM.

Katharine Rainey is the former Director of Educator Development and Support at the Wisconsin Department of Public Instruction.

For more information about this report, please contact Curtis Jones at jones554@uwm.edu or visit www.uwm.edu/sreed.

The Office of Socially Responsible Evaluation in Education conducts rigorous evaluations and research on issues relevant to providing students from all backgrounds with equitable education opportunities.