UWM UNIVERSITY of WISCONSIN MILWAUKEE

SREED Socially Responsible Evaluation in Education

The Role of Interpersonal Bias in the Effectiveness Ratings Assigned to Wisconsin Educators

WEERP Wisconsin Educator Effectiveness Research Partnership

Curtis Jones, Office of Socially Responsible Evaluation in Education

Janurary 2023

This study is the second in a series examining the bias and discrimination affecting Wisconsin teachers of color, as reflected in their performance feedback. In the first study (Jones, Gilman, Reeves, & Rainey, 2021) we found that teachers of color and male teachers receive lower effectiveness ratings across and within schools. The reasons for this were not entirely clear. This second study in the series is designed to isolate any possible racialized or gendered interpersonal bias that might help explain the lower ratings assigned to teachers of color. Are educators of color and male educators viewed as less effective when their administrator is a different gender or race/ethnicity?

Summary Findings

The intersection of teacher and administrator gender and race was a strong predictor of effectiveness ratings. Teachers from different gender and racial backgrounds received different ratings from administrators from different genders and racial backgrounds. We found little evidence that these differences were due to gendered or racialized interpersonal bias though. Where we did find evidence of bias it was limited to specific groups, e.g., female Latinx teachers were rated as more effective by female administrators. Mostly, the results of this study suggest ratings differences reflect different tendencies exhibited by administrators from different backgrounds. This manifested differently for specific combinations of teachers and administrators.

Contents

The Role of Interpersonal Bias in the Effectiveness Ratings Assigned to Wisconsin Educators

This study is the second in a series exploring the systemic and interpersonal bias, and related discrimination, educators of color experience in Wisconsin schools. The focus of our study is performance feedback provided to teachers as part of the Wisconsin Educator Effectiveness (EE) System. 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. In this report, we examine effectiveness ratings to identify evidence of interpersonal bias on perceptions of the effectiveness of Wisconsin educators.

The first study in this series (Jones, Gilman, Reeves, & Rainey, 2021) verified that teachers of color are generally viewed as less effective than their White colleagues (Campbell & Ronfeldt, 2018; Drake et al., 2019). In that study we found administrators view White female teachers as the most effective, with Black and Asian male teachers viewed as the least effective; 89% and 78% of White female teachers are rated as more effective than the average Black and Latinx male teacher, respectively. This was true even when comparing the ratings of teachers with the same credentials and in the same schools.

The first study left open the question as to what types of bias might explain the lower ratings assigned to teachers of color generally and, specifically, male teachers of color. As a group, teachers of color serve more underserved schools and classrooms with more underserved students (Kalogrides et al., 2013), which affects their ability to succeed (Campbell & Ronfeldt, 2018; Steinberg & Sartain, 2020). All of this represents a form of *systemic bias*, in that the contexts where teachers of color work make their jobs harder, which, in turn, make them less likely to excel. It may be discriminatory to rate an educator as less effective because they are working in contexts less organized to promote student and teacher success.

Research is unclear regarding if *interpersonal bias* is also a relevant factor in explaining the lower ratings assigned to teachers of color. For example, in the aforementioned study by Steinberg & Sartain (2020), the authors concluded that interpersonal bias played little to no role in the lower ratings assigned to Chicago teachers of color. However, other research in business (Greenhaus & Parasuraman, 1993; Stauffer & Buckley, 2005; Constantine & Sue, 2007) and in education (Campbell & Ronfeldt, 2018; Drake et al., 2019; Jiang & Sporte, 2016) suggests it does. In a qualitative study of 150 Black educators, teachers 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, in fact, similarly effective (Campbell, 2020). There is evidence that interpersonal bias affects perceptions of male teachers as well, with perceptions that teaching is a more feminine profession and males are less likely to fit into that stereotype (Wind et al., 2019). The current study examines the issue of gendered and racialized interpersonal bias as it manifests in the effectiveness ratings assigned to Wisconsin teachers.

Current Study

In this second study in this series, we use effectiveness ratings of all Wisconsin teachers from 2014-15 to 2019-20. Appendix A presents the characteristics of teachers and administrators involved in the feedback process included in our analyses. Appendix B presents the unadjusted ratings assigned to teachers according to the race and gender of the teacher and administrator. With these data, we examined ratings for evidence of interpersonal bias using two methods.

For the first method, which we call the *fixed administrator* method (Appendix C), we examined the ratings administrators assign to teachers as a function of their race and gender and the race and gender of the teacher they evaluated. This method asks the question "what would the expected difference in ratings assigned by an administrator be when they provide feedback to someone from their same or different racial background and gender?" Administrators are only included in this analysis if they have provided feedback to at least two teachers from different racial or gender backgrounds . Of the 2,447 administrators included in our sample, all but 76 had evaluated more than one teacher, with the average administrator assigning ratings to 20.3 teachers. Two thousand thirty-four administrators provided ratings to both male and female teachers and thus were used to estimate the impact of the gender congruence between administrators and teachers on ratings. Nine hundred ninety-four administrators provided ratings to teachers from at least two different racial backgrounds and thus were used to estimate the impact of the racial congruence between administrators and teachers on ratings. The *fixed administrator* method examines the interaction of the race and gender of both teachers and administrators. By doing this, we can measure the difference in ratings assigned by, for instance, a Black female administrator to a White male teacher, a White female teacher, and a Black female teacher.

For the second method, which we call the *fixed teacher* method (Appendix D), we examined the ratings teachers receive as a function of their race and gender and the race and gender of the administrator who assigned them. The fixed teacher method centers the analysis on teachers, only including teachers who have received more than one evaluation since 2015 from evaluators of different backgrounds. Given that this method looks within teachers to estimate the impact of race and gender congruence between teachers and administrators, it potentially allows us to make causal attributions regarding the differences in ratings assigned to teachers from different backgrounds. The teacher is a constant. Using, what is called, teacher fixed effect modeling, we can predict the difference in ratings assigned to a Black teacher when they receive feedback one year from a White administrator and then when the same teacher, in another year, receives feedback from a Black administrator. The major limitation of this more rigorous method is that it uses a small group of teachers to estimate the impact of gender and race congruence. First, to be included teachers would need to have received feedback more than once. Of the 34,027 teachers included in the population who received ratings, 20,420 (60%) had only received feedback one time and thus are not included in the *fixed teacher* method sample. Of these, 3,661 had received feedback from both a male and female administrator and thus are included in the estimation of the impact of teacher and administrator gender congruence on ratings. Only 1,735 had received feedback from at least two administrators from different racial backgrounds and thus are included in the estimation of the impact of teacher and administrator racial congruence on ratings. However, among these 1,735 teachers, the specific racial groups were very small (Table 8, Appendix D). For instance, only 93 Black teachers had received feedback from both a White and Black administrator. The small group overlap between teacher and administrator racial groups included in these analyses limited our ability to explore the impact of racial group congruence between teachers and administrators on ratings.

Fixed Administrator Results

The *fixed administrator* model (equation 1; Appendix C) was a good predictor of teacher performance ratings, explaining 36.5% of the variance (Table 1). Nearly all model factors were uniquely predictive of teacher ratings. The largest predictor was the administrator who provided it. However, with the model used, the administrator effect includes much of the school effect as well. Measures of teacher experience were also strong predictors of effectiveness ratings. **The intersection of teacher and administrator gender and race was among the strongest predictors of effectiveness ratings.**

Table 1: Fixed administrator model effects predicting FfT teacher ratings

Source	*Type III Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
Corrected Model	1758.508a	1977	0.889	12.24	0
Intercept	409.122	1	409.122	5629.903	0
Teacher/admin – gender/race interaction	**61.24**	**56**	**1.094**	**15.049**	**<.001**
Year	5.14	4	1.285	17.683	<.001
School Type	0.558	3	0.186	2.558	0.053
Fixed administrator effect	937.842	1899	0.494	6.796	<.001
Teacher is new to the school	22.288	1	22.288	306.698	<.001
Average experience of teachers in school	0.314	1	0.314	4.318	0.038
% of students in school - Black	2.747	1	2.747	37.799	<.001
% of students in school - White	0.092	1	0.092	1.26	0.262
% of students in school - low-income	0.688	1	0.688	9.465	0.002
School size	0.748	1	0.748	10.289	0.001
Teacher experience	202.653	1	202.653	2788.692	0
Teacher education	12.251	1	12.251	168.584	<.001
Error	2668.643	36723	0.073		
Total	374771.2	38701			
Corrected Total	4427.151	38700			

R Squared = .397 (*Adjusted R Squared* = .365)

Tables 2 through 4 present a sample of model parameter estimates specific to the teacher/admin - gender/race interaction.[1] Within each teacher group, we look for evidence of both gender and racial bias. The reference groups for these tables are White female teachers evaluated by White female administrators. [2] Thus, the effects (*B*) represent the adjusted ratings scale points higher or lower assigned to a teacher group as compared to the ratings assigned to White female teachers evaluated by White female administrators. For evidence of gender bias, we first look within the specific racial group, i.e., when the teachers and administrators are from the same racial background. Because White administrators provided the most ratings to teachers from all racial groups, we then look for differences in ratings assigned to different genders of teachers by White administrators. For evidence of racial bias, we compare the ratings assigned to teachers by administrators from their same racial background to ratings assigned by White administrators.

Latinx Teachers

Looking at the model results for Latinx teachers (Table 2), all teacher groups were rated as less effective than White female teachers evaluated by a White female administrator, although only of few of these differences were statistically significant. For instance, a Latinx male teacher evaluated by a White male administrator typically received 0.333 scale points lower ratings than a White female teacher rated by a White female administrator.

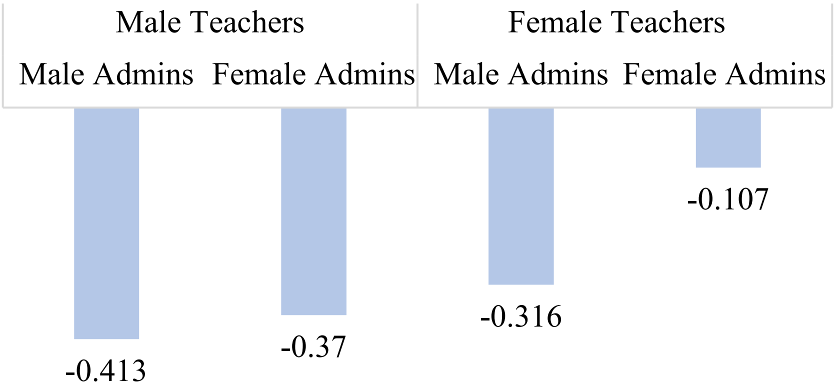
In Figure 1 we present the adjusted difference in ratings for Latinx teachers when the administrator is also Latinx, depending on the gender of both.[3] Female Latinx teachers evaluated by a female Latinx administrator received higher ratings. Also similar, male teachers received nearly the same ratings if they were evaluated by a male or female administrator. **This pattern of results suggests possible evidence of interpersonal gender bias affecting Latinx teachers, but only for female Latinx administrators.**

[1] We had planned to do analyses on teachers from “Other” racial backgrounds but the model only included 20 ratings assigned to teachers from “Other” racial backgrounds by an administrator from “Other” racial backgrounds.

[2] In this study, the reference teacher group for comparison is ratings assigned to White female teachers by a White female administrator. This group was used because of how common it was in our data and because it has close to the highest unadjusted ratings (Appendix B).

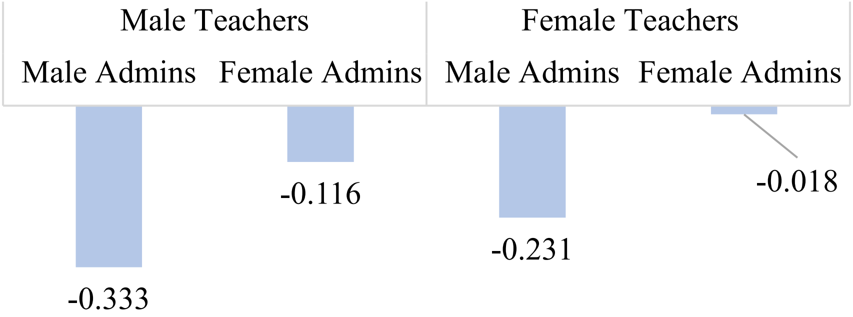
[3] Figure 1 includes 74 ratings assigned to female teachers by a male administrator, 170 ratings assigned to female teachers by a female administrator, 28 ratings assigned to male teachers by a male administrator, and 40 ratings assigned to male teachers by a female administrator.

Figure 1: Fixed administrator model adjusted differences in FfT ratings assigned to Latinx teachers by Latinx administrators by the gender of teachers and administrators



Regarding Latinx teachers evaluated by a White administrator, **both male and female Latinx teachers received lower ratings when evaluated by a White male administrator (Figure 2), suggesting no evidence of gender bias.**[4] Again, ratings were higher regardless of the gender of Latinx teachers and administrators when teachers received ratings from a White administrator than when assigned by a Latinx administrator (Figure 1). **That ratings would be higher when provided by a White administrator again suggests there was no evidence of interpersonal racial bias influencing the ratings of Latinx teachers.**

Figure 2: Fixed administrator model adjusted differences in FfT ratings assigned to Latinx teachers by White administrators by the gender of teachers and administrators



[4] Figure 2 includes 209 ratings assigned to female teachers by a male administrator, 396 ratings assigned to female teachers by a female administrator, 68 ratings assigned to male teachers by a male administrator, and 102 ratings assigned to male teachers by a female administrator.

Table 2: Fixed administrator model adjusted ratings assigned to Latinx teachers

Gender		Race/ethnicity		*B*	*Std. Error*	*t*	*Sig.*
Teacher	Admin	Teacher	Admin				
male	male	Latinx	White	-0.333	0.121	-2.742	0.006
male	male	Latinx	Black	-0.177	0.132	-1.342	0.18
male	male	Latinx	Latinx	-0.413	0.289	-1.427	0.153
male	female	Latinx	White	-0.116	0.028	-4.15	<.001
male	female	Latinx	Black	-0.324	0.114	-2.827	0.005
male	female	Latinx	Latinx	-0.37	0.215	-1.715	0.086
female	male	Latinx	White	-0.231	0.12	-1.936	0.053
female	male	Latinx	Black	-0.104	0.131	-0.792	0.428
female	male	Latinx	Latinx	-0.316	0.287	-1.101	0.271
female	female	Latinx	White	-0.018	0.016	-1.073	0.283
female	female	Latinx	Black	-0.218	0.109	-1.992	0.046
female	female	Latinx	Latinx	-0.107	0.213	-0.503	0.615

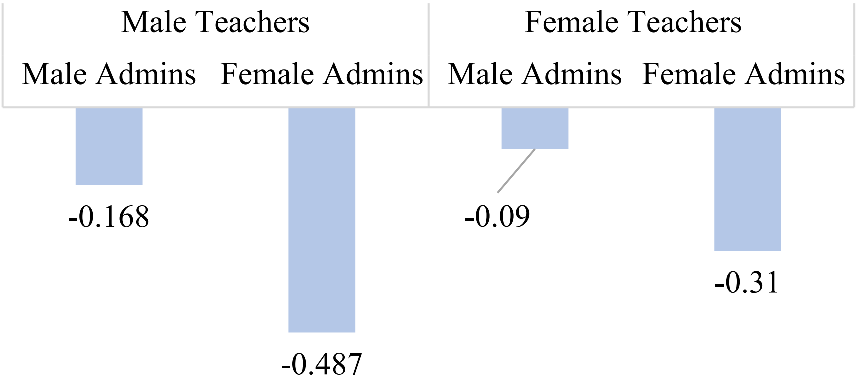
Black Teachers

Looking at the model results for Black teachers (Table 3), all teacher groups were rated as less effective than White female teachers evaluated by a White female administrator, with many of these statistically significant. For instance, a Black male teacher evaluated by a White male administrator typically received 0.411 scale points lower ratings than a White female teacher rated by a White female administrator.

In Figure 3 we present the adjusted difference in ratings for Black teachers when the administrator is also Black, depending on the gender of both.[5] Contrary to the results for Latinx teachers, **Black male teachers evaluated by a Black administrator received much lower ratings, regardless of the administrator's gender.**

[5] Figure 3 includes 101 ratings assigned to female teachers by a male administrator, 262 ratings assigned to female teachers by a female administrator, 44 ratings assigned to male teachers by a male administrator, and 93 ratings assigned to male teachers by a female administrator.

Figure 3: Fixed administrator model adjusted differences in FfT ratings assigned to Black teachers by Black administrators by the gender of teachers and administrators



Regarding Black teachers evaluated by a White administrator, again similar to what was found with Latinx teachers (Figure 2), **both male and female Black teachers received lower ratings when evaluated by a White male administrator (Figure 4), suggesting no evidence of gender bias.**[6]

In an interesting dynamic, compared to ratings provided by White administrators, ratings were 0.243 points higher for Black male teachers when provided by a Black male administrator (-0.411 compared to -0.168) but 0.261 points lower when provided by a Black female administrator (-0.226 compared to -0.487).

Again compared to ratings provided by White administrators, ratings were 0.205 more effective for Black female teachers when assigned by Black male administrator (-0.295 compared to -0.090). However, Black female teachers were typically rated 0.193 points less effective (-0.117 compared to -0.310 points) when rated by a Black female administrator.

[6] Figure 4 includes 536 ratings assigned to female teachers by a male administrator, 1,092 ratings assigned to female teachers by a female administrator, 260 ratings assigned to male teachers by a male administrator, and 348 ratings assigned to male teachers by a female administrator.

Figure 4: Fixed administrator model adjusted differences in FfT ratings assigned to Black teachers by White administrators by the gender of teachers and administrators

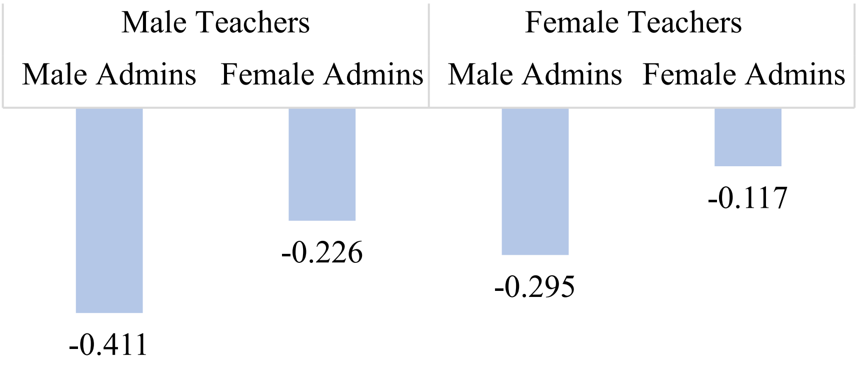


Table 3: Fixed administrator model adjusted ratings assigned to Black teachers

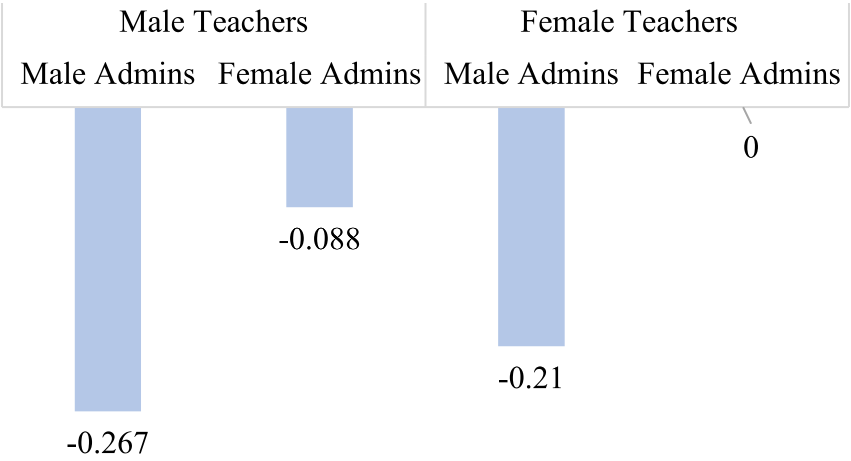
Gender		Race/ethnicity		*B*	*Std. Error*	*t*	*Sig.*
Teacher	Admin	Teacher	Admin				
male	male	Black	White	-0.411	0.123	-3.34	<.001
male	male	Black	Black	-0.168	0.124	-1.358	0.174
male	male	Black	Latinx	-0.571	0.307	-1.862	0.063
male	female	Black	White	-0.226	0.032	-6.987	<.001
male	female	Black	Black	-0.487	0.106	-4.579	<.001
male	female	Black	Latinx	-0.361	0.228	-1.585	0.113
female	male	Black	White	-0.295	0.121	-2.436	0.015
female	male	Black	Black	-0.09	0.122	-0.74	0.459
female	male	Black	Latinx	-0.314	0.297	-1.056	0.291
female	female	Black	White	-0.117	0.021	-5.624	<.001
female	female	Black	Black	-0.31	0.104	-2.995	0.003
female	female	Black	Latinx	-0.201	0.217	-0.927	0.354

White Teachers

Looking at the model results for White teachers (Table 4), all teacher groups were rated as less effective than White female teachers evaluated by a White female administrator. For instance, a White male teacher evaluated by a White male administrator typically received 0.267 scale points lower ratings than a White female teacher rated by a White female administrator.

In Figure 5 we present the adjusted difference in ratings for White teachers when the administrator is also White, depending on the gender of both.[7] Unique compared to the pattern of results for Black or Latinx teachers, female administrators assigned higher ratings to both male and female White teachers. **The pattern of these results does not suggest that interpersonal gender bias affects the ratings assigned to White teachers.**

Figure 5: Fixed administrator model adjusted differences in FfT ratings assigned to White teachers by White administrators by the gender of teachers and administrators



[7] Figure 5 includes 8,406 ratings assigned to female teachers by a male administrator, 8,358 ratings assigned to female teachers by a female administrator, 3,664 ratings assigned to male teachers by a male administrator, and 2,164 ratings assigned to male teachers by a female administrator.

Table 4: Fixed administrator model adjusted ratings assigned to White teachers

Gender		Race/ethnicity		*B*	*Std. Error*	*t*	*Sig.*
Teacher	Admin	Teacher	Admin				
male	male	White	Black	-0.17	0.12	-1.421	0.155
male	male	White	Latinx	-0.455	0.287	-1.586	0.113
male	male	White	White	-0.267	0.118	-2.257	0.024
male	female	White	Black	-0.368	0.103	-3.576	<.001
male	female	White	Latinx	-0.299	0.212	-1.412	0.158
male	female	White	White	-0.088	0.006	-14.797	<.001
female	male	White	Black	-0.077	0.119	-0.653	0.514
female	male	White	Latinx	-0.292	0.285	-1.025	0.305
female	male	White	White	-0.21	0.118	-1.774	0.076
female	female	White	Black	-0.226	0.102	-2.204	0.028
female	female	White	Latinx	-0.129	0.212	-0.607	0.544
female	female	White	White	0a	.	.	.

Fixed Teacher Results

The fixed teacher model (equation 2; Appendix D) explained 10.6% of the overall variance in effectiveness ratings (11.1% across teachers and 10.3% within teachers). Teacher job experience was the strongest predictor of ratings, with teachers receiving higher ratings as they became more experienced. As outlined in Appendix D, due to the small sample of teachers, the fixed teacher method limits our ability to examine the ratings assigned to many combinations of teachers and administrator groups. What we were able to examine regarding the impact of gender and race/ethnicity of teachers and administrators is summarized below.

Gender Bias

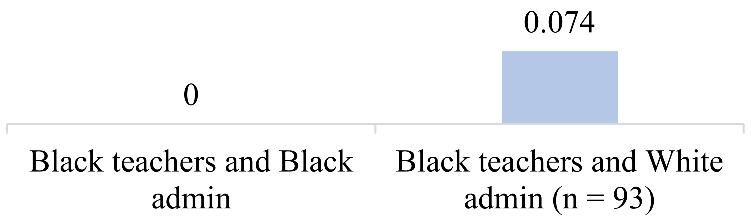
We first found, what seemed like, evidence that male teachers were rated as more effective when they received feedback from a male evaluator ($p = .007$). Specifically, male teachers received 0.025 scale points higher ratings when rated by a male administrator. However, female teachers were also rated as 0.015 scale points more effective when their ratings were assigned by a male administrator ($p = .006$). Thus, it is more accurate to say that male administrators rate teacher performance slightly more favorably regardless of the teacher's gender. This is also consistent with the results of the fixed administrator model. The fixed teacher model results suggest that, across racial groups, lower ratings of male teachers were not explained by the intersection of the genders of the administrator and teacher.

Race/Ethnicity Bias

Regarding interpersonal race/ethnicity bias, due to the dominance of White teachers and administrators in Wisconsin (Jones, 2019), the data did not allow us to reliably estimate the impact of an administrator's race on the effectiveness ratings for many combinations of teacher and administrator racial/ethnic groups. For instance, only 18 Black teachers had documented ratings from both Black and Latinx administrators. Further, we were also not able to analyze the impact of racial congruence for Native American, Asian, or Pacific Islander teachers. Too few teachers from these groups received multiple ratings in the Wisconsin EE system to include (Appendix D). Even when we collapsed these into one "Other" group there were still too few teachers.

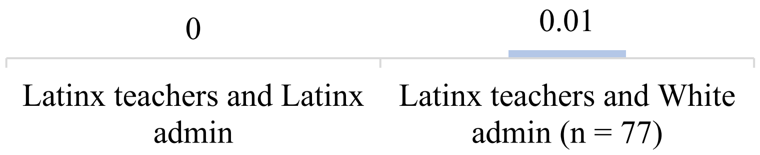
Regarding Black teachers, we are only able to report the differences in ratings assigned when receiving them from both a White and Black administrator (Figure 6). The other groups were too small to reliably estimate. Black teachers as a group received higher ratings from White administrators than they did when they were evaluated by Black administrators (*p* = .068). The magnitude of the difference translates to 0.074 scale points. Again though, this overall effect seems to blur the more complex effect presented in Figures 3 and 4. We found that Black male administrators rated Black male teachers more positively than White male administrators, and that White female administrators rated Black female teachers more positively than Black female administrators. Without accounting for the interaction between race and gender, the overall effect presenting in fixed teacher analysis seems to be an oversimplification of how Black teacher effectiveness ratings are influenced by the race and gender of the administrator.

Figure 6: Fixed teacher model adjusted differences in FfT ratings assigned to Black teachers by a White and a Black administrator



Regarding Latinx teachers, we are only able to report the difference in ratings when assigned by both a White and Latinx administrator (Figure 7). There was no difference between the ratings Latinx teachers received from a White or Latinx administrator. This is in contrast to the fixed administrator results presented in Figures 1 and 2, which suggest White administrators rated Latinx teachers more positively than Latinx administrators. Again, these results show that the utility of the fixed teacher analysis for understanding the impact of race and gender on Latinx teacher ratings is limited.

Figure 7: Fixed teacher model adjusted differences in FfT ratings assigned to Latinx teachers by a White and a Latinx administrator



Summary and Discussion

We used six years of statewide effectiveness ratings data assigned to teachers as part of the Wisconsin EE System to examine evidence of possible gender and racial interpersonal bias. We conducted two sets of analyses with these data. First, we used a fixed administrator model. Only administrators who had assigned ratings to more than one teacher were included in this model. With this method, the race or gender of the teachers rated by a single administrator had to vary. Next, we used a fixed teacher model. Only teachers who had received ratings more than one time were included in this model. With this method, the race and gender of the administrators who rated a teacher had to vary.

The results from the fixed administrator model yielded several interesting findings about how administrator and teacher race and gender intersect to help explain the ratings assigned to teachers. After accounting for differences in teacher and school characteristics, along with the fixed effect of the administrator providing ratings, we found evidence that the intersection of teacher and administrator gender and race was a strong predictor of effectiveness ratings. This effect manifested differently for different combinations of teacher and administrator racial groups. For example, we found that female Latinx teachers were rated as more effective by female administrators. However, both female and male administrators rated male Latinx teachers as less effective. Thus, for Latinx teachers, we found evidence of possible bias exhibited by female administrators but not male administrators.

We also measured gendered tendencies of higher or lower ratings assigned to teachers from different racial backgrounds. Male administrators, compared to female administrators, rated both male and female Latinx and White teachers as less effective. This was also true regarding ratings assigned by White male administrators to male and female Black teachers. However, the opposite was true regarding Black female administrators who rated Black female teachers as less effective. While none of these results suggest gender bias, in that each administrator group assigned consistent ratings to both male and female teachers from a specific racial background, they do still suggest gendered tendencies for rating teachers.

We also found little evidence that racial bias explained much about the ratings assigned to teachers. We again found ratings tendencies for administrators of different racial backgrounds and genders. White male administrators assigned Black teachers lower ratings than Black male administrators regardless of teacher gender. However, White female administrators assigned Black teachers higher ratings than Black female administrators. White administrators, regardless of gender, also assigned Latinx teachers higher ratings than Latinx administrators respectively. That different administrator demographic groups had different tendencies for rating teachers is still problematic. Two teachers in the same district might expect different ratings based on the background and rating tendencies of their administrator. **Although not related to gendered or racialized bias, this finding suggests a teacher's rating is partially determined on the characteristics of their administrator rather than just their performance as a teacher.**

The results from the fixed teacher model for measuring evidence of interpersonal bias was severely limited by sample size constraints. Even considering the nearly 50,000 effectiveness rating records, too few teachers met the conditions required by the method to make many useful and reliable comparisons. This prevented us from analyzing the data in a way that respects the complexity of how race and gender influence effectiveness ratings. The rarity of teachers of color in Wisconsin, and especially experienced teachers of color, makes this type of analysis difficult (Jones, 2019). Wisconsin teachers typically receive performance feedback every four years. Thus, the teachers included in this study would have had to remain as a teacher for at least that long. We know from our previous research that few educators of color remain teachers that long (Jones, 2019).

Our study is consistent with previous research that suggests the racial match of the teacher and administrator do not explain the ratings assigned to teachers once school, teacher, and student characteristics are accounted for (Steinberg & Sartain, 2020). However, our study does suggest research on the influence of interpersonal bias on effectiveness ratings is oversimplifying the issue by analyzing rating across genders and racial groups. How race in Wisconsin is related to effectiveness ratings is dependent on the intersection of the race and gender of teachers and administrators. Future research on interpersonal bias in effectiveness systems in other contexts should be more nuanced than has been done previously. Given the underrepresentation of educators of color in our education system, research that accounts for the complexity of the interplay between race and gender will require very large datasets, such as the statewide data presented in this paper. Qualitative research could also be useful to understanding how the race and gender of teachers and administrators intersect to influence measures of teacher effectiveness. Although the current study does not provide any evidence that racialized or gendered interpersonal bias is impacting the ratings assigned to teachers, it likely still occurs. Understanding the conditions where is happens will provide guidance for preventing it.

Appendix A – Sample

This study includes all Framework for Teaching (FfT) ratings assigned to classroom teachers participating in the Wisconsin Educator Effectiveness (EE) System from 2016 to 2020 who could be linked to a specific administrator who provided them performance feedback. During this time, 49,546 effectiveness ratings were assigned to 34,027 teachers. EE system data also document the administrator who assigned ratings to each educator. 2,380 unique administrators assigned ratings to teachers. Through linking both sets of data to WiseStaff data managed by the Wisconsin Department of Public Instruction, we were able to collect additional information about both educators and administrators, including their education, experience, race/ethnicity, and gender. Teacher characteristics like experience, education, and school varied across the time of the study. Gender and race are fixed in the data and are reported below (Table 5).

Table 5: Teacher and administrator characteristics

	N	%
Administrators		
Male	1164	48.9
Female	1216	51.1
White	2014	84.6
Black	266	11.2
Latinx	73	3.1
Other	27	1.1
Teachers		
Male	8671	25.5
Female	25356	74.5
White	31269	91.9
Black	1084	3.2
Latinx	1059	3.1
Other	615	1.8

School characteristics included in sample

We were also able to link EE ratings data with the district and school the teacher served. This allowed us to include demographic information about the school including the percent of students from different racial/ethnic backgrounds and the percent eligible for free or reduced lunch. Teachers reflected in the EE data served 262 districts and 1340 schools. 774 schools were elementary, 230 were middle, and 296 were high schools. Forty schools served students across grade bands. The teaching force in a typical school had 11.7 years of teaching experience.

Table 6: Descriptive statistics of 1,340 schools in study

	Mean	SD
% with an IEP	16.0	6.9
% Econ Disadv	48.8	26.0
% Black	11.2	22.5
% White	68.1	29.2
School size	421.7	324.7
Overall teacher experience	11.7	3.1

Appendix B – Unadjusted ratings assigned to teachers according to the race and gender of the teacher and administrator

Table 7: Unadjusted ratings assigned to teachers according to the race and gender of the teacher and administrator

Teacher Race	Admin Race	Teacher Gender	Admin Gender	Mean	N	Std. Deviation
White	White	Male	Male	3.08	6266	0.30
White	White	Male	Female	3.03	3707	0.34
White	White	Female	Male	3.14	14632	0.29
White	White	Female	Female	3.13	14358	0.32
White	Black	Male	Male	2.87	386	0.34
White	Black	Male	Female	2.82	538	0.45
White	Black	Female	Male	2.99	866	0.34
White	Black	Female	Female	2.96	1778	0.39
White	Latinx	Male	Male	2.91	111	0.40
White	Latinx	Male	Female	2.87	154	0.51
White	Latinx	Female	Male	3.12	313	0.40
White	Latinx	Female	Female	3.03	581	0.40
White	Other	Male	Male	3.10	94	0.38
White	Other	Male	Female	3.04	56	0.30
White	Other	Female	Male	3.04	236	0.42
White	Other	Female	Female	3.10	259	0.28
Black	White	Male	Male	2.82	84	0.36
Black	White	Male	Female	2.77	96	0.34
Black	White	Female	Male	2.92	170	0.39
Black	White	Female	Female	2.85	249	0.47
Black	Black	Male	Male	2.77	82	0.38
Black	Black	Male	Female	2.58	150	0.52
Black	Black	Female	Male	2.90	170	0.40
Black	Black	Female	Female	2.81	459	0.45
Black	Latinx	Male	Male	2.75	8	0.47
Black	Latinx	Male	Female	2.72	15	0.50
Black	Latinx	Female	Male	2.97	13	0.47
Black	Latinx	Female	Female	2.92	47	0.45
Black	Other	Male	Male	*	*	*
Black	Other	Male	Female	*	*	*
Black	Other	Female	Male	2.70	15	0.52

Cont on next page…

Teacher Race	Admin Race	Teacher Gender	Admin Gender	Mean	N	Std. Deviation
Black	Other	Female	Female	2.78	12	0.68
Latinx	White	Male	Male	2.91	137	0.39
Latinx	White	Male	Female	2.92	121	0.44
Latinx	White	Female	Male	3.07	319	0.29
Latinx	White	Female	Female	3.03	381	0.33
Latinx	Black	Male	Male	2.84	27	0.32
Latinx	Black	Male	Female	2.84	39	0.40
Latinx	Black	Female	Male	3.02	30	0.42
Latinx	Black	Female	Female	2.94	65	0.40
Latinx	Latinx	Male	Male	2.88	41	0.28
Latinx	Latinx	Male	Female	2.70	63	0.43
Latinx	Latinx	Female	Male	3.04	114	0.30
Latinx	Latinx	Female	Female	2.97	246	0.39
Latinx	Other	Male	Male	*	*	*
Latinx	Other	Male	Female	3.19	5	0.19
Latinx	Other	Female	Male	*	*	*
Latinx	Other	Female	Female	3.35	7	0.39
Other	White	Male	Male	2.99	86	0.38
Other	White	Male	Female	2.93	92	0.37
Other	White	Female	Male	3.04	234	0.30
Other	White	Female	Female	3.05	268	0.34
Other	Black	Male	Male	2.74	10	0.40
Other	Black	Male	Female	2.68	27	0.45
Other	Black	Female	Male	2.87	35	0.41
Other	Black	Female	Female	2.93	73	0.40
Other	Latinx	Male	Male	2.53	4	0.51
Other	Latinx	Male	Female	*	*	*
Other	Latinx	Female	Male	3.13	13	0.45
Other	Latinx	Female	Female	2.99	24	0.49
Other	Other	Male	Male	2.73	7	0.47
Other	Other	Male	Female	2.63	6	0.53
Other	Other	Female	Male	2.77	14	0.58
Other	Other	Female	Female	2.99	9	0.29

* Results are suppressed because sample includes fewer than five ratings.

Appendix C – Fixed administrator analytic approach

We used fixed administrator modeling in SPSS 28 to isolate the relationship between racial/ethnic and gender congruence between teachers and their administrator and the ratings assigned to teachers. Fixed administrator effects estimate the difference in ratings assigned by an administrator when they assign ratings to at least two teachers. In the case of the current study, we were interested in examining if administrators who provided ratings to teachers from different backgrounds, in relation to their background, rated teachers differently. For instance, did White female administrators rate the effectiveness of White female, Black female, or White male teachers differently. Evidence of gendered or racialized bias would manifest if White female administrators rated teachers with a different gender or from a different racial group differently. To explore this, we used the model below:

$$
\begin{aligned}
Y_{it} = \beta_0 &+ \beta_1(Teacher\ experience_{ita}) + \beta_2(Overall\ teacher\ experience_{ita}) \\
&+ \beta_3(Teacher\ new\ to\ school_{ita}) + \beta_4(Teacher\ education_{ita}) \\
&+ \beta_5(Teacher\ race\ \times\ Admin\ race\ \times\ Teacher\ gender\ \times Admin\ gender_{ita}\ ) \\
&+ \beta_6(schoolFRlunch_{ita}) + \beta_7(\%schoolWhite_{ita}) + \beta_8(\%schoolBlack_{ita}) \\
&+ \beta_9(School\ size_{ita}) + \beta_{10}(School\ type_{ita}) + \sum_{y=1}^{y-1} \beta_{11.m}\, Year_{ia} + \sum_{a=1}^{a-1} \beta_{12.m}\, Administrator_{it} \\
&+ \epsilon_{ita}
\end{aligned}
\tag{1}
$$

where Y_i is the overall FfT rating for the i^{th} person, in t^{th} time, within a^{th} administator; β_0 is the intercept; β_1 is the effect of teacher experience; β_2 is the effect of the overall experience of teachers a school; β_3 is the effect of a teacher being new to a school; β_4 is the effect of teacher education; β_5 is the effect of the interaction between admin and teacher gender and race; β_6 is the effect of the % of school eligible for F/R lunch; β_7 is the effect of the school % who are White; β_8 is the effect of the school % who are Black; β_9 is the effect of school size; β_{10} is the effect of school type; β_{11} is the effect of year the rating was assigned, β_{12} is fixed administrator effect, and ϵ_i is error term for the i^{th} in person in t^{th} time within a^{th} administrator.

Appendix D – Fixed teacher analytic approach

We used fixed teacher modeling with the XTREG function in STATA 12.0 to isolate the relationship between racial/ethnic and gender congruence between teachers and their administrator and the ratings assigned to teachers. Fixed teacher effects estimate the difference in ratings assigned to a teacher when they receive ratings from at least two administrators. Because teachers are compared to themselves, fixed teacher effects potentially allow us to make causal attributions about factors that change exogenous to the teacher. In the case of the current study, we were interested in examining if teachers who had received at least two evaluations were rated different by each evaluator according to the evaluator's background. For instance, were Black male teachers rated as more effective when they were rated by a Black female administrator than by a White female administrator? To explore this question, we used the model below:

$$\begin{aligned} Y_{it} = \beta_0 &+ \beta_1(Teacher\ experience_{it}) + \beta_2(Teacher\ new\ to\ school_{it}) + \beta_3(Admin\ education_{it}) \\ &+ \beta_4(Admin\ gender_{it}) \\ &+ \beta_5(Admin\ race_{it}) + \beta_6(Admin\ experience_{it}) + \beta_7(Admin\ role\ (AP\ or\ principal)_{it}) \\ &+\ \beta_8(Teacher\ gender\ \times Admin\ gender_{it}) + \beta_9(Teacher\ Race\ \times Admin\ Race_{it}) \\ &+ \beta_{10}(schoolFRlunch_{it}) + \beta_{11}(\%schoolWhite_{it}) + \beta_{12}(\%schoolBlack_{it}) + \beta_8(School\ size_{it}) \\ &+ \sum_{i=1}^{i-1} \beta_{13.i}\, Year_t + \sum_{t=1}^{t-1} \beta_{14.t}\, Teacher_i + \epsilon_{it} \end{aligned} \tag{2}$$

where Y_i is the overall FfT rating for the i^{th} person in t^{th} time, β_0 is the intercept; β_1 is the effect of teacher experience; β_2 is the effect of a teacher being new to a school; β_3 is the effect of a admin education; β_4 is the effect of a admin gender; β_5 is the effect of a admin race; β_6 is the effect of a admin experience; β_7 is the effect of a admin role; β_8 is the effect of the interaction between teacher and admin gender; β_9 is the effect the interaction between teacher and admin race; β_{10} is the effect of the school % F/R lunch status; β_{11} is the effect of the school % who are White; β_{12} is the effect of the school % who have an IEP; $\beta_{13.i}$ is the effect of year the rating was assigned, $\beta_{14.t}$ is fixed teacher effect; and ϵ_i is error term for the i^{th} in person in t^{th} time.

In this method, our ability to estimate the difference in ratings assigned to teachers in difference demographic conditions requires differences in the demographic conditions of administrators across assessments. We considered including the intersections of teacher and administrator race and gender in the model (β_5 from equation 1), but the sample of specific race and gender interaction groups was too small to include in the model. Because of this, we were limited to including the interactions terms of teacher and administrator gender (β_8 in equation 2) and race (β_9 in equation 2) separately. Even without interacting gender and race, sample size remained limiting. For gender, there were 3,661 teachers who had received ratings one year by a female administrator and in another year by a male administrator. 975 of these teachers were male and 2,686 were female. These 3,661 teachers are the sample from which we isolate the impact of gender congruence on ratings. For race/ethnicity, the relevant sample was much smaller. There were 1,799 teachers who received ratings one year by an administrator from one racial/ethnic group and then in another year by an administrator from a different racial/ethnic group. Table 8 breaks down the racial/ethnic background of teachers and administrators represented among these 1,799 teachers. From these we can see that White teachers most commonly received ratings from administrators from difference racial/ethnic groups. For instance, 1,454 White teachers (5% of all White teachers with ratings and 11.7% of White teachers with ratings in more than one year) received ratings from administrators from two different racial/ethnic groups. Specially, 840 received ratings from both a White and Black administrator. 145 Black teachers (13% of all Black teachers with ratings and 33.6% of Black teachers with ratings in more than one year) received ratings from administrators from two different racial/ethnic groups. Specially, 93 received ratings from both White and Black evaluators. 152 Latinx teachers (14% of all Latinx teachers with ratings and 32.7% of Latinx teachers with ratings in more than one year) received ratings from administrators from two different racial/ethnic groups. Specifically, 77 received ratings from both White and Latinx evaluators. With these small group sizes, the fixed teacher analytic approach should result in a reasonably precise measure of the impact of gender congruence but less precise for race/ethnicity congruence between teachers and administrators.

Table 8: Numbers of teachers from different racial/ethnic groups evaluated by multiple administrators from different racial/ethnic groups.

	White and Black Admins	White and Latinx Admins	White and Other Admins	Black and Latinx Admins	Black and Other Admins	Latinx and Other Admins	Total Teachers with Ratings from at least Two Admins from Different Racial Groups	Teachers with Ratings in More than one Year	Percent of Teachers Rated at Least Twice who were Rated by Admins from Different Racial Groups
White Teachers	840	316	179	91	21	7	1,454	12,444	11.7%
Black Teachers	93	19	7	18	7	1	145	432	33.6%
Latinx Teachers	52	77	3	17	2	1	152	465	32.7%
Other Teachers	25	12	6	4	1	0	48	266	18.0%

Appendix E – Educator Effectiveness Rubric

The Danielson Framework for Teaching (FfT) (2013) measures educator effectiveness across 22 components and four domains. 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.

For the purposes of this study, we created an overall FfT 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. Our use of overall ratings is for research purposes only and does not reflect the typical practice of how educators receive performance feedback in Wisconsin. Across all teachers, the average FfT ratings was 3.08 with a standard deviation of 0.33.

In the first study of this series, we also included ratings assigned in schools using the Strong (2002) framework. We did not include these Stronge ratings in the current study because so few teachers and administrators of color use this framework in Wisconsin.

References

Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). *Teacher Demographics and Evaluation: A Descriptive Study in a Large Urban District*. Available online at https://files.eric.ed.gov/fulltext/ED569346.pdf

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. Available online at 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.* Available online at https://nces.ed.gov/pubs97/97535l.pdf

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. Available online at https://edtrust.org/wp-content/uploads/2014/09/ThroughOurEyes.pdf

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. Available online at https://consortium.uchicago.edu/sites/default/files/2018-10/Teacher%20Evaluation%20in%20Chicago-Jan2016-Consortium.pdf

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

Jones, C.J., Gilman, L., Reeves, M., & Rainey, K. (2021). *Evidence of Discrimination and Bias in the Effectiveness Ratings Assigned to Wisconsin Educators of Color.* Available online at https://uwm.edu/sreed/wp-content/uploads/sites/502/2021/06/Bias-and-discrimination-reflected-in-effectiveness-ratings.pdf

Kini, T., & Podolsky, A. (2016). Doe*s Teaching Experience Increase Teacher Effectiveness? A Review of the Research.* Palo Alto: Learning Policy Institute. Available online at https://learningpolicyinstitute.org/sites/default/files/product-files/Teaching_Experience_Report_June_2016.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability. [Monograph.]*. RAND Corporation. Available online at https://www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf

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. Jou*rnal 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.

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.

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.

Thanks to William Cannon, Jessica Cordaro, Carl Frederick, Thomas Johnson, Katherine Rainey, and Alexandria Sedar for their helpful feedback on drafts on this report.

SREED™
Socially Responsible Evaluation in Education

UWM®
UNIVERSITY of WISCONSIN
MILWAUKEE