Abstract Title Page

Title: An Observer Like Me: The Effect of Demographic Congruence Between Teachers and Raters on Classroom Observation Scores

Author: Olivia L. Chi, Harvard University

Background

U.S. states and districts have come to rely on teacher evaluations – and in particular, classroom observations – as a key lever for teacher accountability. However, recent studies suggest that classroom observation scores may be influenced by factors that are beyond teachers' control (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). Evidence also suggests that administrators' complex work environments influence how administrators rate teachers in high-stakes contexts (Cohen & Goldhaber, 2016; Grissom & Loeb, 2017; Kraft & Gilmour, 2017; Qi et al., 2018).

Furthermore, a growing body of research suggests that demographic congruence between teachers and principals may play an important role in teacher outcomes, such as hiring, turnover, and job satisfaction (Bartanen & Grissom, 2019; Grissom & Keiser, 2011; Husain, Matsa, & Miller, 2018), and demographic congruence can be an important factor in the observation and feedback cycle (Kraft & Christian, 2019). Taken together, recent work raises concerns about whether and to what extent teachers' and administrators' demographic characteristics influence classroom observation scores.

Research Questions

1. What is the effect of demographic congruence between teachers and observers on teachers' classroom observation ratings?
2. Are the effects of demographic congruence mediated by the sharing of other attributes, such as education history or teaching assignment history?

Setting/Population

I use administrative data from a large district in the southeastern United States from 2013-18. My analytic sample includes 93,975 classroom observations from 38,262 teacher-years from 12,490 unique teachers. These observations were conducted by 2,319 observer-years from 672 unique observers. This sample has been restricted to the two race subgroups for which I have a substantial sample size: Black and White teachers who are observed by Black and White administrators. See Table 1 for descriptive statistics.

Context/Practice

In this district, each teacher is required to be observed 2-3 times per school year. During classroom observations, the observer (a principal or assistant principal) evaluates teachers on the state's evaluation rubric.

Research Design/Analysis

To address RQ1, I exploit the availability of multiple rounds of observation scores per teacher during each school year and the within teacher-year variation in the demographic characteristics of the classroom observers. To identify the impact of demographic matching between teachers and observers, I estimate models that include both teacher-by-year fixed effects and observer-by-round-by-year fixed effects:

$$y_{ijkt} = \beta_0 + \beta_1 Match\_C_{ijkt} + Z_{ijkt} + \delta_{it} + \pi_{jkt} + u_{ijkt} \quad (1)$$

y_{ijkt} is the observation score belonging to teacher i, rated by observer j, in observation round k in school year t. $Match\_C_{ijkt}$ is an indicator that equals 1 if teacher i and observer j share the same characteristic of interest C (e.g., race). Z_{ijkt} is a vector of observation-level covariates (functions of time length, starting hour, and month).

δ_{it} represents teacher-by-year fixed effects, which control for unobserved teacher quality and other characteristics that are invariant within year t. The inclusion of teacher-by-year fixed effects implies that the identifying variation comes from teacher-years in which the teacher is observed by at least one rater who shares the characteristic of interest and at least one rater who does not. By including δ_{it}, I compare a teacher's observation score to the other observation scores she received in the same school year t.

π_{jkt} represents observer-by-round-by-year fixed effects. Including π_{jkt} accounts for unobserved and observed differences in rater characteristics across raters, observation rounds, and time. They also control for shocks that are common across all the observations conducted by observer j in observation round k in school year t. Standard errors are two-way clustered at the teacher-level and observer-level.

β_1 represents the average effect of a teacher and the classroom observer sharing the characteristic C. The identifying assumption is that, among the observations that a teacher receives in the same school year, selection into having an observer who shares the characteristic C is uncorrelated with unobserved determinants of observation scores.

To address RQ2, I first generate variables that capture other commonalities between teachers and observers, such as whether they ever taught the same grade/content and whether they attended the same university. I re-estimate equation 1, including these variables in the right-hand side. Changes in the coefficients on the race or gender congruence indicators provide evidence of the extent to which these measures of commonalities act as mediators.

Results

I find that teachers, on average, experience a small increase in observation scores from sharing race (0.03 SD) or gender (0.02 SD) with their observers (Table 2). For comparison, these magnitudes are about 10% and 8%, respectively, of the average within-teacher returns to experience after one year of teaching.

Using another specification (see Appendix A for details), I also examine how the magnitude of race and gender gaps change when administrators, who belong to the underperforming group, conduct classroom observations. I find that the Black-White observation score gap is smaller by 0.06 SD when teachers are rated by Black observers, as compared to when teachers are rated by White observers (Table 3). Similarly, the male-female gap is smaller by 0.05 SD when teachers are rated by male observers, as compared to when teachers are rated by female observers. These magnitudes are non-trivial, representing roughly one-third of the unconditional Black-White score gap and one-quarter of the unconditional male-female gap, respectively.

Furthermore, while I do find that the extent of professional familiarity between teachers and observers (as measured by years of working in the same school) is significantly related to observation scores, I do not find that any of my included relationship measures of commonalities or familiarity mediate the race and gender congruence effects (Table 4). The race and gender

dynamics between teachers and raters appear to exist separately from these relationship characteristics.

Conclusions

As with prior research on demographic congruence, the mechanisms at work are unclear. Additional research is needed, perhaps in the form of field experiments designed to test possible mechanisms. Even though the underlying mechanisms are unclear, the results raise fairness concerns for teachers whose demographics are not reflected by any of their administrators. These results implore those who use observation scores in decision-making to carefully consider the circumstances and context under which the scores were generated.

References

Bartanen, B., & Grissom, J. A. (2019). *School principal race and the hiring and retention of racially diverse teachers* (EdWorkingPaper No. 19–59). Retrieved from http://edworkingpapers.com/ai19-59

Campbell, S. L., & Ronfeldt, M. (2018). Observational Evaluation of Teachers: Measuring More Than We Bargained for? *American Educational Research Journal*, *55*(6), 1233–1267. https://doi.org/10.3102/0002831218776216

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387. https://doi.org/10.3102/0013189X16659442

Fairlie, R. W., Hoffmann, F., & Oreopoulos, P. (2014). A Community College Instructor Like Me: Race and Ethnicity Interactions in the Classroom. *American Economic Review*, *104*(8), 2567–2591. https://doi.org/10.1257/aer.104.8.2567

Grissom, J. A., & Keiser, L. R. (2011). A supervisor like me: Race, representation, and the satisfaction and turnover decisions of public sector employees: Race, Representation, and the Satisfaction and Turnover Decisions of Public Sector Employees. *Journal of Policy Analysis and Management*, *30*(3), 557–580. https://doi.org/10.1002/pam.20579

Grissom, J. A., & Loeb, S. (2017). Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low- and High-Stakes Environments. *Education Finance and Policy*, *12*(3), 369–395. https://doi.org/10.1162/EDFP_a_00210

Husain, A. N., Matsa, D. A., & Miller, A. R. (2018). *Do Male Workers Prefer Male Leaders? An Analysis of Principals' Effects on Teacher Retention* (No. NBER WP No. 25263). National Bureau of Economic Research.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting *The Widget Effect*: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Educational Researcher*, *46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

Qi, Y., Bell, C. A., Jones, N. D., Lewis, J. M., Witherspoon, M. W., & Redash, A. (2018). *Administrators' uses of teacher observation protocol in different rating contexts* (Research Report No. ETS RR-18-18; pp. 1–19). Retrieved from http://doi.wiley.com/10.1002/ets2.12205

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. https://doi.org/10.3102/0162373715616249

Table 1: Descriptive Statistics for Analytic Sample

	Analytic Sample				Race Congruence FE Sample				Gender Congruence FE Sample			
	Teachers		Observers		Teachers		Observers		Teachers		Observers	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
White	0.87	0.34	0.71	0.45	0.84	0.37	0.57	0.50	0.88	0.33	0.72	0.45
Black	0.13	0.34	0.29	0.45	0.16	0.37	0.43	0.50	0.12	0.33	0.28	0.45
Female	0.81	0.40	0.60	0.49	0.80	0.40	0.59	0.49	0.82	0.38	0.51	0.50
Num. of Observations	2.51	0.50			2.65	0.48			2.65	0.48		
Has both White & Black Observers	0.21	0.41			1.00	0.00			0.38	0.48		
Has both Male & Female Observers	0.31	0.46			0.57	0.50			1.00	0.00		
Experience	12.32	8.91			11.17	8.82			11.12	8.84		
Has Tenure	0.48	0.50			0.38	0.48			0.40	0.49		
Is Principal			0.35	0.48			0.32	0.47			0.33	0.47
Is Assistant Principal			0.65	0.48			0.68	0.47			0.67	0.47
Num. of Observations Conducted			44.25	15.55			42.71	14.95			43.96	15.43
N (person-years)	38,262		2,319		7,985		1,136		12,017		1,457	

Note: The sample includes data from 12,490 unique teachers and 672 unique observers in school years 2013-14 through 2017-18. The race (gender) congruence fixed effects sample refers to the person-years that conttribute to the identifying variation for estimating the effect of race (gender) congruence on observation scores.

Table 2. Demographic Congruence

	Observation Score	
	(1)	(2)
Race Match	0.031*	
	(0.012)	
Gender Match		0.024*
		(0.010)
Teacher-year FE	Y	Y
Observation controls	Y	Y
Observer-round-year FE	Y	Y
Teacher-years	38262	38262
Observer-years	2319	2319
Observations	93975	93975

Notes: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001. Two-way clustered standard errors (teacher-level and observer-level) are in parentheses. Estimated model is in equation 1. Observation controls include indicators for observation month, indicators for starting hour, and a quadratic function of the time length.

Table 3. Changes in Gaps

	Observation Score	
	(1)	(2)
Black Match	0.061*	
	(0.024)	
Male Match		0.047*
		(0.020)
Teacher-year FE	Y	Y
Observation controls	Y	Y
Observer-round-year FE	Y	Y
Teacher-years	38262	38262
Observer-years	2319	2319
Observations	93975	93975

Notes: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001. Two-way clustered standard errors (teacher-level and observer-level) are in parentheses. Estimated model is in equation A1. Observation controls include indicators for observation month, indicators for starting hour, and a quadratic function of the time length.

Table 4: Exploring Mediators

	Outcome: Observation Score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Race Match	0.031*	0.030*	0.029*	0.030*	0.030*	0.030*	0.030*
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Gender Match	0.024*	0.023*	0.023*	0.023*	0.023*	0.023*	0.023*
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Teaching Assignment Match	-0.017	-0.018					-0.019
	(0.015)	(0.015)					(0.015)
Attended Same University	0.021		0.020				0.023
	(0.017)		(0.017)				(0.017)
Attended University in Same State	-0.019			-0.019			-0.024
	(0.017)			(0.017)			(0.017)
Yrs. Same School	0.017*				0.017+		0.016+
	(0.009)				(0.009)		(0.009)
Yrs. Same School^2	-0.001+				-0.001+		-0.001+
	(0.001)				(0.001)		(0.001)
Yrs. Same School-Team	0.016					0.015	0.014
	(0.020)					(0.020)	(0.020)
Yrs. Same School-Team^2	-0.002					-0.002	-0.002
	(0.003)					(0.003)	(0.003)
Observation controls	Y	Y	Y	Y	Y	Y	Y
Teacher-year FE	Y	Y	Y	Y	Y	Y	Y
Observer-round-year FE	Y	Y	Y	Y	Y	Y	Y
Teacher-years	38262	38262	38262	38262	38262	38262	38262
Observer-years	2319	2319	2319	2319	2319	2319	2319
Observations	93975	93975	93975	93975	93975	93975	93975

Notes: + p<0.10 * p<0.05 ** p<0.01 *** p<0.001. Two-way clustered standard errors (teacher-level and observer-level) are in parentheses. Observation controls include indicators for observation month, indicators for starting hour, and a quadratic function of the time length. In column 1, each group of coefficients separated by a solid line are estimates from a separate regression. Each column of 2-7 reports results from a single regression.

Appendix A

Estimating Changes in Race and Gender Gaps

To examine how race and gender gaps change when administrators, who belong to the underperforming group, conduct classroom observations, I adopt a strategy similar to that used by Fairlie, Hoffmann, and Oreopoulos (2014) who examine performance gaps between underrepresented minority and white community college students when taught by underrepresented minority instructors. Specifically, to examine race gaps, I fit:

$$y_{ijkt} = \lambda_0 + \lambda_1 BlackMatch_{ij} + \delta_{it} + \pi_{jkt} + u, \qquad \text{(A1)}$$

where $BlackMatch_{ij}$ is an indicator variable that equals 1 if both teacher i and observer j identify as Black. The coefficient λ_1 provides an estimate of the change in Black teachers' scores (relative to White teachers' scores) when the observer is also Black, as compared to Black teachers' relative scores when the observer is White. In other words, λ_1 provides an estimate of whether the Black-White gap in observation scores is larger or smaller when observations are conducted by Black observers, as compared to White observers. λ_1 is positive if Black teachers' relative scores are higher when observed by a Black administrator, relative to that when observed by a White administrator. In the context of the existing gap, a positive value of λ_1 would indicate that the gap between Black and White teachers is smaller under Black observers.

To examine where the male-female gap in observation scores is larger or smaller under male observers, as compared to female observers, I replace the variable $BlackMatch_{ij}$ with $MaleMatch_{ij}$ in equation A1. Analogously, $MaleMatch_{ij}$ equals 1 if both teacher i and observer j are males, and the coefficient on $MaleMatch_{ij}$ is positive if male teachers' relative scores are higher when observed by male administrators.

Here, λ_1 could be biased if there exists some factor that: (1) coincides with being rated by a Black (male) observer; (2) relates to the conditional outcomes, and (3) exists for Black (male) teachers but not other teachers. One such threat stems from differential sorting, which occurs if, for example, highly motivated Black (male) teachers sort to Black (male) classroom observers, while highly motivated White (female) teachers do not.