Online Appendix

for Rangel and Shi's "First Impressions Matter: Evidence from Elementary-School Teachers"

A. Tables and Figures

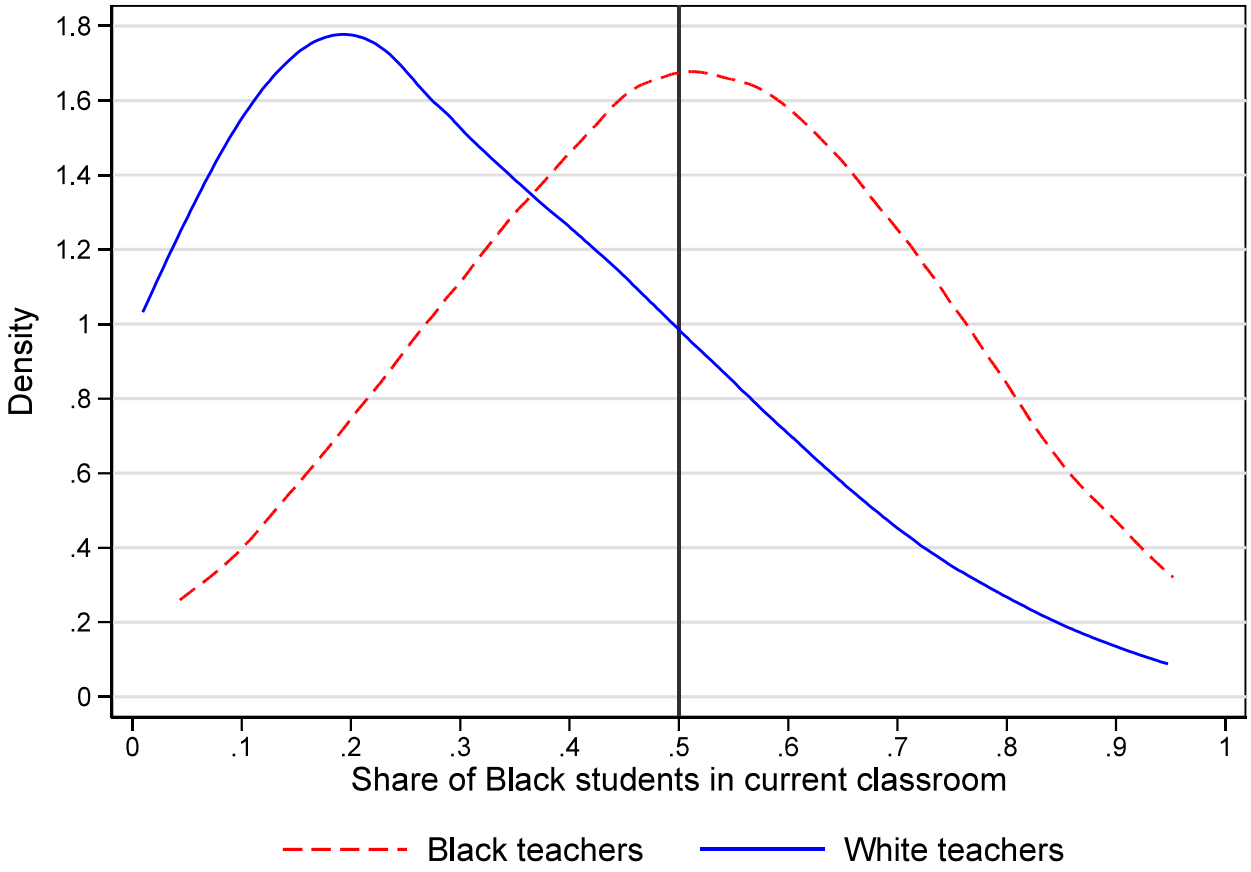
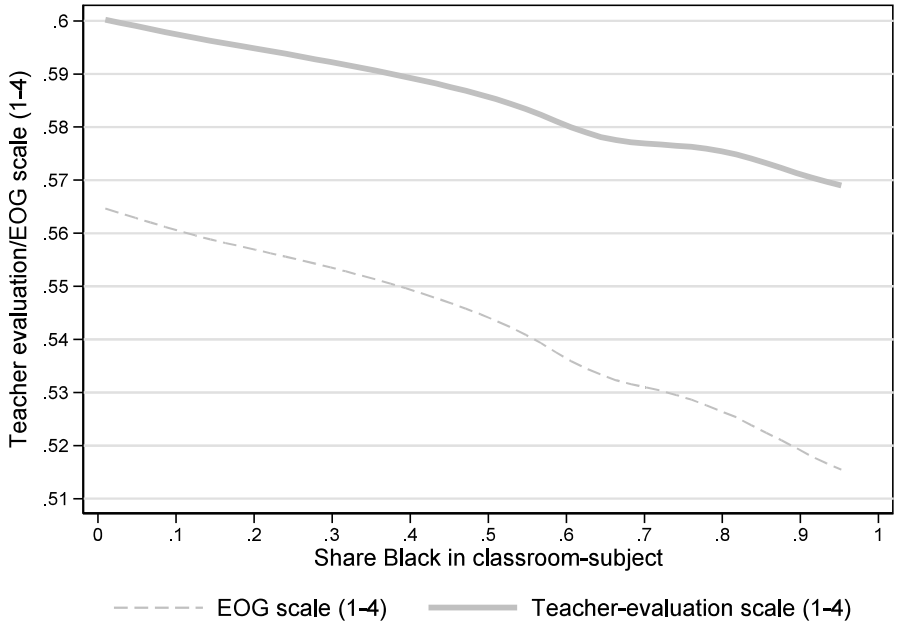
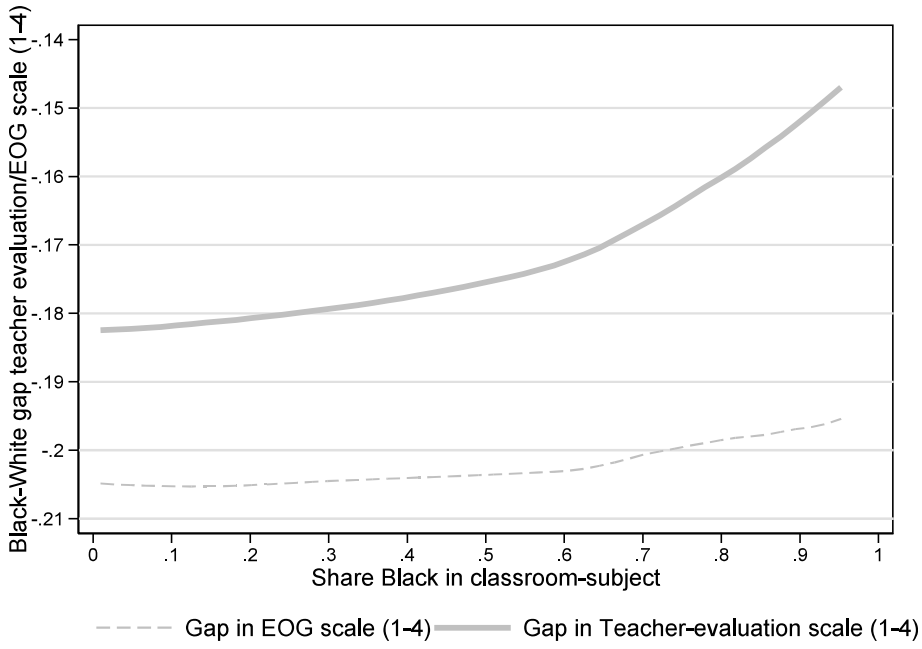


Figure OA1: Classroom-level racial composition by teacher race



Panel A: Black Students' Teacher Evaluation and Standardized Test Score levels



Panel B: Black-White Gap in Teacher Evaluation and Standardized Test Score levels

Figure OA2: Classroom-level objective and subjective measures versus racial composition

Table OA1: Variance decomposition within and across classrooms

	Within variation share (1)	Between variation share (2)
Panel A: Our working sample		
a. Teacher evaluation levels (1-4) - Math	0.86	0.14
b. EOG achievement levels (1-4) - Math	0.74	0.26
Ratio a/b	1.16	0.54
Difference a-b	0.12	-0.12
c. Teacher evaluation levels (1-4) - Reading	0.86	0.14
d. EOG achievement levels (1-4) - Reading	0.79	0.21
Ratio c/d	1.09	0.67
Difference c-d	0.07	-0.07
Observations Math	101,995	101,995
Observations Reading	101,067	101,067
Panel B: Elder and Zhou (2020) - 3rd graders, ECLS-K 2011 [Table 4, column 4]		
a. Math rating	0.92	0.08
b. IRT test scores - Math	0.69	0.31
Ratio a/b	1.44	0.26
Difference a-b	0.30	-0.30
c. Reading rating	0.91	0.09
d. IRT test scores - Reading	0.68	0.32
Ratio c/d	1.34	0.28
Difference c-d	0.23	-0.23
Observations	NA	NA

Table OA2: Initial Classroom Characteristics and Novice Teacher Attributes - Novice-teacher allocation

	Share Black (1)	Average EOG, Black (2)	White-Black Gap (3)
Panel A: Raw analysis			
Female teacher	−0.009	−0.026	0.023
	(0.015)	(0.032)	(0.044)
Black teacher	0.160***	−0.075*	0.075
	(0.018)	(0.040)	(0.058)
Teacher education: BA only	0.038***	−0.067*	−0.100**
	(0.014)	(0.034)	(0.040)
Teacher Licensed	−0.045***	0.036	0.022
	(0.010)	(0.024)	(0.029)
Observations	1,907	1,907	1,907
Panel B: Conditional on initial school fixed effects			
Female teacher	−0.003	0.011	0.006
	(0.009)	(0.037)	(0.054)
Black teacher	0.001	−0.051	0.130*
	(0.012)	(0.047)	(0.070)
Teacher education: BA only	0.006	0.019	−0.010
	(0.008)	(0.043)	(0.051)
Teacher Licensed	0.003	0.022	−0.011
	(0.006)	(0.030)	(0.039)
Observations	1,575	1,575	1,575

Notes: This table shows the correspondence between novice teacher observable characteristics and those of their initial classrooms. Sample after school-fixed effects is reduced due to singletons. *** p<0.01, ** p<0.05, * p<0.1.

Table OA3: Robustness to Alternative Specifications of EOG Test Scores

	Common support Trimmed tails	Common support Discretized	EOG FE
	(1)	(2)	(3)
Panel A: Teacher Evaluation Scale (1-4)			
Black	−0.060***	−0.057***	−0.057***
	(0.004)	(0.007)	(0.004)
EOG test score	0.682***	0.683***	
	(0.004)	(0.013)	
Panel B: Teacher Evaluation (Proficient=1)			
Black	−0.024***	−0.022***	−0.023***
	(0.003)	(0.004)	(0.003)
EOG test score	0.375***	0.391***	
	(0.003)	(0.009)	
Panel C: Teacher Evaluation (Above class mean=1)			
Black	−0.031***	−0.034***	−0.028***
	(0.003)	(0.004)	(0.003)
EOG test score	0.369***	0.394***	
	(0.003)	(0.008)	
Observations	187,536	39,422	203,042

Notes: This table uses alternative specifications of EOG test scores to examine whether there is sufficient within-classroom overlap of Black and White test score distributions. Column 1 trims the tails of performance distributions so that the sample ranges from the maximum of the lowest Black and White scorers up to the minimum of the top scorers by race. Column 2 only keeps observations for which a Black student has a White classmate with the same EOG score for a given subject. Column 3 includes EOG fixed effects. All standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1.

Table OA4: Robustness of Teacher Evaluation Bias to Inclusion of Socio-demographic, Academic, and Behavioral Covariates – Teacher Evaluation Scale (1-4)

	(1)	(2)	(3)	(4)	(5)	(6)
Black	−0.060***	−0.053***	−0.071***	−0.065***	−0.059***	−0.054***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)
EOG test scores	0.682***	0.544***	0.682***	0.679***	0.687***	0.657***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)	(0.006)
Male	−0.036***	−0.028***	−0.035***	−0.029***	−0.031***	−0.025***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
One year younger than mode	0.038	0.028	0.030	0.031	0.040	0.035
	(0.039)	(0.039)	(0.039)	(0.040)	(0.047)	(0.048)
One year older than mode	−0.052***	−0.037***	−0.048***	−0.047***	−0.046***	−0.045***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Two years older than mode	−0.164***	−0.111***	−0.150***	−0.146***	−0.145***	−0.140***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.012)	(0.012)
Three+ years older than mode	−0.163***	−0.112*	−0.138**	−0.136**	−0.120	−0.109
	(0.057)	(0.057)	(0.058)	(0.058)	(0.073)	(0.072)
Additional controls included						
Days absent indicators		YES	YES	YES	YES	YES
Lag student evaluation indicators		YES				
Suspension days indicators				YES	YES	YES
Free-reading indicators						YES
Homework indicators						YES
Sample restrictions (due to information availability)						
Restricted to 2008-2013			YES	YES		
Restricted to 2008-2012					YES	YES
Observations	203,062	203,062	195,471	195,471	118,579	118,579

Notes: All standard errors are clustered at the teacher level. EOG test scores are included as z-scores centered at state-mandated proficiency cutoff and as a fourth-order polynomial function. Reported coefficient on EOG test scores is the marginal effect evaluated at the proficiency cutoff. Additional controls are accounted for semi-parametrically with a set of indicator functions. Days of suspension are included separately for each infraction. *** p<0.01, ** p<0.05, * p<0.1

Table OA5: Teacher Subject-Mastery Evaluation Scale (1-4), Student Race, and Standardized Test Scores (EOG) - by subject/grade/teacher experience strata

	By subject		By grade		By teacher experience		
	Math	Reading	Grade 4	Grade 5	1 year	2 years	3 years
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Black	−0.039***	−0.079***	−0.055***	−0.063***	−0.062***	−0.057***	−0.058***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.008)	(0.009)
EOG test score	0.714***	0.657***	0.701***	0.666***	0.674***	0.682***	0.698***
	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.007)	(0.009)
Male	−0.026***	−0.051***	−0.043***	−0.031***	−0.033***	−0.050***	−0.024***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.006)
One year younger than mode	0.017	0.062	0.074	0.008	0.019	0.113*	−0.025
	(0.045)	(0.047)	(0.063)	(0.048)	(0.058)	(0.058)	(0.091)
One year older than mode	−0.036***	−0.065***	−0.050***	−0.054***	−0.058***	−0.050***	−0.045***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Two years older than mode	−0.131***	−0.193***	−0.165***	−0.163***	−0.167***	−0.155***	−0.171***
	(0.011)	(0.011)	(0.014)	(0.012)	(0.014)	(0.015)	(0.017)
Three+ years older than mode	−0.091	−0.233***	−0.243***	−0.067	−0.249***	−0.180*	0.012
	(0.064)	(0.068)	(0.066)	(0.095)	(0.069)	(0.092)	(0.143)
Observations	101,995	101,067	96,831	106,231	87,749	67,419	47,894
Classroom-subjects	3,750	3,744	3,928	3,566	3,332	2,444	1,718
Teachers	2,195	2,191	1,210	1,160	1,644	1,207	846

Notes: All standard errors are clustered at the teacher's unique ID level. EOG test scores are included as z-scores centered at grade-subject state-mandated proficiency cutoff and as a fourth-order polynomial function. Reported coefficient on EOG test scores is the marginal effect evaluated at the proficiency cutoff. Demographic controls include indicators for gender and age relative to the within-grade modal age. *** p<0.01, ** p<0.05, * p<0.1

Table OA6: Teacher Subject-Mastery Evaluation Scale (1-4), Student Race, and Standardized Test Scores (EOG) - by child gender and age strata

	By gender		By age	
	Girls	Boys	At or below mode	Above mode
	[1]	[2]	[3]	[4]
Black	−0.068***	−0.054***	−0.067***	−0.044***
	(0.006)	(0.006)	(0.005)	(0.007)
EOG test score	0.674***	0.683***	0.673***	0.696***
	(0.006)	(0.005)	(0.005)	(0.006)
Male	-	-	−0.032***	−0.045***
			(0.004)	(0.006)
One year younger than mode	−0.025	0.110	0.031	-
	(0.054)	(0.068)	(0.041)	
One year older than mode	−0.046***	−0.058***	-	0.120**
	(0.005)	(0.005)		(0.061)
Two years older than mode	−0.133***	−0.184***	-	0.017
	(0.014)	(0.012)		(0.061)
Three+ years older than mode	−0.082	−0.221***	-	-
	(0.099)	(0.077)		
Observations	89,170	90,828	114,513	64,984
Classroom-subjects	6,313	6,520	6,688	5,774
Teachers	1,943	1,987	2,025	1,851

Notes: All standard errors are clustered at the teacher's unique ID level. EOG test scores are included as z-scores centered at grade-subject state-mandated proficiency cutoff and as a fourth-order polynomial function. Reported coefficient on EOG test scores is the marginal effect evaluated at the proficiency cutoff. Demographic controls include indicators for gender and age relative to the within-grade modal age. *** p<0.01, ** p<0.05, * p<0.1

Table OA7: Initial Classroom (IC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale) – strata by current student gender

	Full sample	Sample with IC information			
	(1)	(2)	(3)	(4)	(5)
Panel A: Boys only					
Black	−0.054***	−0.055***	−0.053***	−0.052***	−0.006
	(0.006)	(0.007)	(0.007)	(0.007)	(0.016)
Black × Share of Black in IC			0.013	0.008	0.011
			(0.035)	(0.035)	(0.035)
Black × White-Black score gap in IC				−0.030***	
				(0.011)	
Black × 1{White-Black score gap in IC > 0}					−0.054***
					(0.017)
Observations	90,828	71,259	71,259	71,259	71,259
Classroom-subjects	6,520	5,345	5,345	5,345	5,345
Teachers	1,987	1,770	1,770	1,770	1,770
Panel B: Girls only					
Black	−0.068***	−0.071***	−0.071***	−0.070***	−0.012
	(0.006)	(0.007)	(0.007)	(0.007)	(0.016)
Black × Share of Black in IC			0.054	0.050	0.055
			(0.036)	(0.035)	(0.036)
Black × White-Black score gap in IC				−0.039***	
				(0.012)	
Black × 1{White-Black score gap in IC > 0}					−0.069***
					(0.017)
Observations	89,170	69,515	69,515	69,515	69,515
Classroom-subjects	6,313	5,164	5,164	5,164	5,164
Teachers	1,943	1,726	1,726	1,726	1,726

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), gender, and age indicators. The initial classroom (IC) sample restricts to observations with racial mix (at least one student of each race) and measured White-Black gaps in lagged test scores. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. All interactions with other non-White racial/ethnic groups are also included in the model so that coefficients juxtaposes between Black and White students. *** p<0.01, ** p<0.05, * p<0.1

Table OA8: Robustness of Effect of Initial Classroom (IC) Conditions to Interactions with Teacher Attributes

	Base model (1)	+Black × IC student count (2)	+Black × Tch. covariates (3)	+Black × initial school FE (4)
Panel A: White-Black score gap in IC				
Black × White-Black score gap in IC	−0.033***	−0.031***	−0.031***	−0.020**
	(0.009)	(0.009)	(0.009)	(0.009)
Interacted controls				
Black × Share Black in IC	0.023	0.022	0.001	−0.097
	(0.026)	(0.027)	(0.027)	(0.066)
Black × Share Non-White Non-Black in IC		0.037	0.024	−0.101
		(0.037)	(0.037)	(0.072)
Black × # students in IC		−0.002	−0.002	0.002
		(0.002)	(0.001)	(0.002)
Black × Female teacher			−0.022	−0.009
			(0.016)	(0.017)
Black × Black teacher			0.015	0.008
			(0.027)	(0.025)
Black × Teacher has MA			−0.051***	−0.057***
			(0.015)	(0.017)
Black × Teacher licensed in NC			−0.026**	−0.026**
			(0.011)	(0.013)
Black × Teacher experience			−0.007	−0.008
			(0.006)	(0.006)
Panel B: Sign of White-Black score gap in IC				
Black × 1{White-Black score gap in IC > 0}	−0.056***	−0.053***	−0.052***	−0.041***
	(0.013)	(0.013)	(0.013)	(0.013)
Interacted controls				
Black × Share Black in IC	0.027	0.025	0.004	−0.101
	(0.026)	(0.027)	(0.027)	(0.066)
Black × Share Non-White Non-Black in IC		0.032	0.020	−0.109
		(0.037)	(0.037)	(0.072)
Black × # students in IC		−0.002	−0.001	0.002
		(0.002)	(0.002)	(0.002)
Black × Female teacher			−0.021	−0.009
			(0.016)	(0.017)
Black × Black teacher			0.015	0.008
			(0.027)	(0.025)
Black × Teacher has MA			−0.051***	−0.056***
			(0.016)	(0.018)
Black × Teacher licensed in NC			−0.026**	−0.027**
			(0.011)	(0.013)
Black × Teacher experience			−0.006	−0.007
			(0.006)	(0.006)
Observations	156,291	156,291	156,285	156,285
Classroom-subjects	6,011	6,011	6,011	6,011
Techers	1,907	1,907	1,907	1,907

Notes: All models are estimated using the set of controls listed in Col 5 and 6 of Table 3, which include EOG test scores, gender, age and month of birth indicators, as well as interactions of Black and IC share of Black students. The initial classroom sample restricts to observations with racial composition information. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. All interactions with other non-White racial/ethnic groups are also included in the model so that coefficients juxtaposes between Black and White students. Observations in Column 3 exclude singletons. *** p<0.01, ** p<0.05, * p<0.1.

Table OA9: Initial (IC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale): Permanence of effects

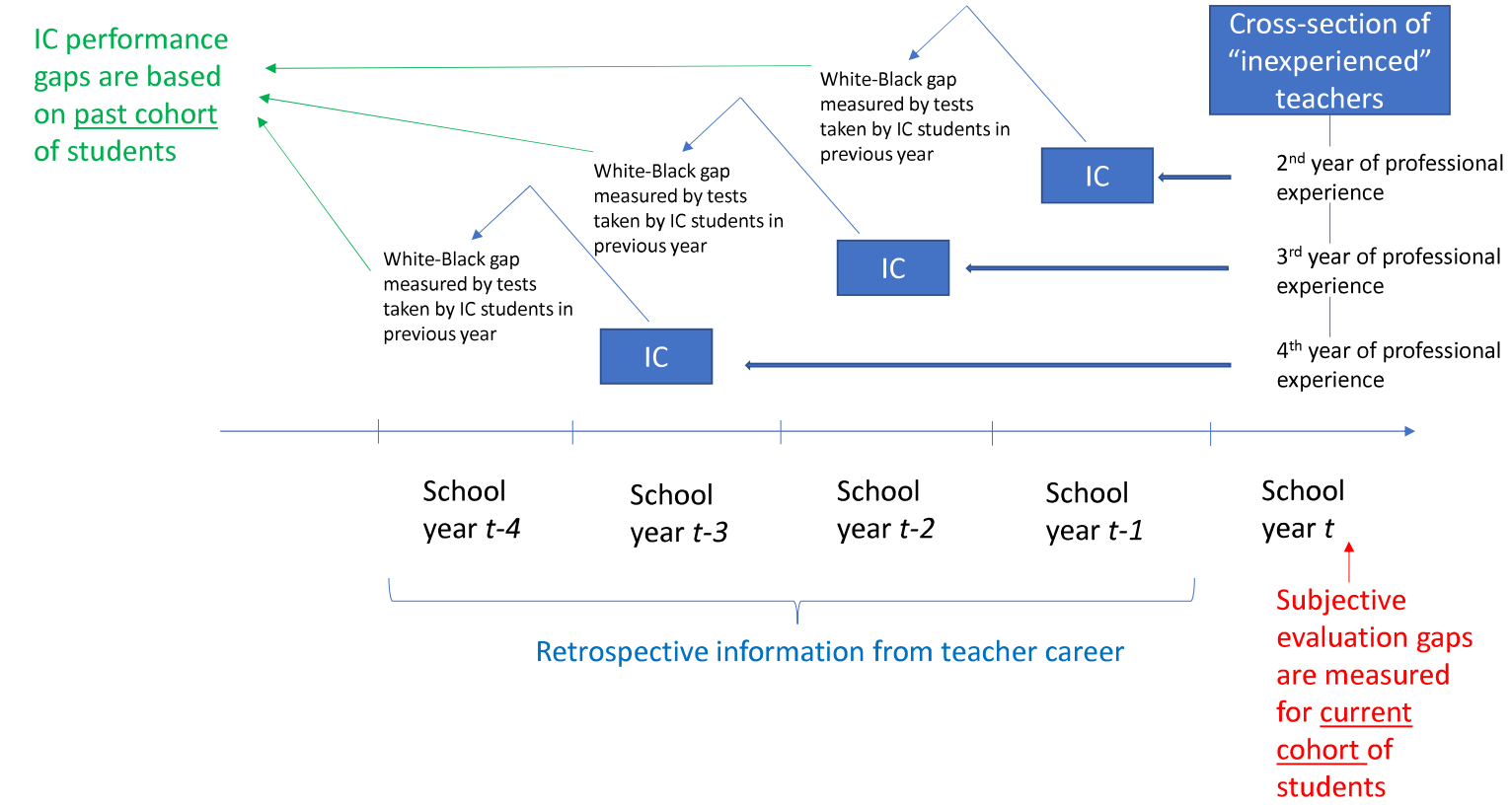
	Teachers in 2nd professional year (1)	Teachers in 3rd professional year (2)	Teachers in 4th professional year (3)	Teachers in 5th to 7th professional year (4)
Panel A: White-Black score gap in IC				
Black × White-Black score gap in IC	-0.025^{**}	-0.023	-0.049^{***}	-0.026^{***}
	(0.012)	(0.014)	(0.015)	(0.009)
Panel B: Sign of White-Black score gap in IC				
Black × $1\{$White-Black score gap in IC $> 0\}$	-0.025	-0.085^{***}	-0.057^{**}	-0.030^{**}
	(0.017)	(0.023)	(0.026)	(0.015)
Observations	67,485	51,782	37,024	70,284
Classroom-subjects	2,687	1,951	1,373	2,501
Teachers	1,457	1,060	737	723

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), gender, and age indicators. The initial classroom (IC) sample restricts to observations with racial mix (at least one student of each race) and measured White-Black gaps in lagged test scores. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. All interactions with other non-White racial/ethnic groups are also included in the model so that coefficients juxtaposes between Black and White students. Controls are also included at the level of the interaction with Black indicator variables. These include the IC size, the share of non-Black and non-White students in the IC, a female teacher indicator, a Black teacher indicator, a teacher with MA degree indicator, a licensed teacher indicator, and teacher experience in years. *** p<0.01, ** p<0.05, * p<0.1

Table OA10: Initial (IC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale): Salience of extremes

	(1)	(2)	(3)
Panel A: Shares of White student population in IC relative to anchoring Black student performances			
Black × Share of White students outscoring highest scoring Black student in IC	−0.010		0.002
	(0.020)		(0.020)
Black × Share of White students outscoring lowest scoring Black student in IC		−0.102***	−0.104***
		(0.033)	(0.034)
Panel B: Indicators for performance differences at tails			
Black × 1{Highest scoring White outscores highest scoring Black student in IC}	−0.011		−0.010
	(0.012)		(0.012)
Black × 1{Lowest scoring White outscores lowest scoring Black student in IC}		−0.021*	−0.021*
		(0.012)	(0.012)
Panel C: Size of performance differences at tails			
Black × [Highest White score - highest Black score in IC]	−0.011*		−0.009
	(0.006)		(0.006)
Black × [Lowest White score - lowest Black score in IC]		−0.009	−0.008
		(0.006)	(0.006)
Observations	156,291	156,291	156,291
Classroom-subjects	6,011	6,011	6,011
Teachers	1,907	1,907	1,907

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), gender, and age indicators. The initial classroom (IC) sample restricts to observations with racial mix (at least one student of each race) and measured White-Black gaps in lagged test scores. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. All interactions with other non-White racial/ethnic groups are also included in the model so that coefficients juxtaposes between Black and White students. Controls are also included at the level of the interaction with Black indicator variables. These include the IC size, the share of non-Black and non-White students in the IC, a female teacher indicator, a Black teacher indicator, a teacher with MA degree indicator, a licensed teacher indicator, teacher experience in years and initial school fixed-effects. *** p<0.01, ** p<0.05, * p<0.1



IC = Initial classroom assignment

Diagram OA1: Data Organization

B. More on reference bias

We consider the possibility that teachers may rely on relative comparisons across students for their subjective assessments. In fact, it is precisely because of this concern that our estimates of racial differences rely on within-classroom variation and are anchored on blind-scored standardized tests. Based on our reading of Elder and Zhou (2021) and additional examinations of our data presented below, we conclude that our empirical approach sufficiently approximates their strategies to correct for reference bias. We first emphasize that within our study context, there are reasons to believe relative comparisons are accompanied by notions of absolute performance clearly communicated to teachers by education authorities. The North Carolina Department of Public Instruction collects teacher evaluation data as part of the regular testing procedure (distinct from the Early Childhood Longitudinal Study IES-sponsored survey questionnaires utilized in Elder and Zhou's analyses). As we pointed out in the text, teachers were told explicitly to focus their evaluation of student mastery only on the tested subject in state-administered exams and were given a scale that is identical to the one used for standardized exams.

To complement this view, we present additional evidence that our context seems less prone to concerns about reference bias relative to the one faced by Elder and Zhou (2021) after undertaking some of the same analyses. Variance decompositions in Table OA1 indicate that differences in the contribution of between-classroom variation in subjective and objective measures of performance are closer to each other in our sample than in Elder and Zhou (2021), Table 4. Figure OA1 then shows that classroom racial composition and subjective and objective measures of performance follow each other more closely in our study context than in Elder and Zhou (2021), Figure 1, Panels A, B, D and E. While both share a negative slope in our sample, Elder and Zhou (2021) report a positive slope between teacher evaluations and school-level share of Black students, while showing a negative slope between objective test scores and the share of Black students. Most importantly, we show that the Black-White gap in both subjective and objective evaluations in our sample follow a parallel pattern across different classroom compositions. These are indications that our sample does not suffer substantially from the potential reference biases raised by the authors in the context of ECLS-K.

Ultimately, Elder and Zhou (2021) derive Black-White gaps in non-cognitive skills that address reference bias by assuming that variation in objective measures such as tests of cognitive skill are informative about latent distributions in non-cognitive skills. In their second approach, they generate a measure of school-level reference bias by taking the difference between the average objective, or blind-scored, measure of cognitive skill and the average subjective measure of cognitive skill. In doing so they make an assumption that we share in our paper, which is that blind-graded standardized test scores are free of reference bias. Deviations in subjective teacher ratings from these objective measures of cognitive skills at the school level are then added as school-specific reference bias to observed non-cognitive skills to get an adjusted Black-White non-cognitive gap.

Finally, we differ in an important way from Elder and Zhou (2021). Instead of relying on within-school variation (a design that also underpins their third suggested approach), we use classroom fixed effects. This means our racial gaps rely on the very localized context of within-class variation alone, and net out any classroom-specific biases such as the propensity to give everyone uniformly higher ratings. This furthermore includes class-specific reference bias shared across students.