Teachers of Color, Culturally Responsive Teaching, and Student Outcomes: Experimental Evidence from the Random Assignment of Teachers to Classes

David Blazar
University of Maryland

December 2021

Abstract

There is broad consensus across academic disciplines that access to same-race/ethnicity teachers is a critical resource for supporting the educational experiences and outcomes of Black, Hispanic, and other students of color. While theoretical and qualitative lines of inquiry further describe a set of teacher mindsets and practices aligned to "culturally responsive teaching" as likely mechanisms for these effects, to date there is no causal evidence on this topic. In experimental data where upper-elementary teachers were randomly assigned to classes, I find large effects upwards of 0.45 standard deviations of teachers of color on the short- and longer-term social-emotional, academic, and behavioral outcomes of their students. These average effects are explained in part by teachers' growth mindset beliefs that student intelligence is malleable rather than fixed, interpersonal relationships with students and families, time spent planning for and differentiating instruction for individual students' needs, and the extent to which teachers lead well-organized classrooms in which student (mis)behavior is addressed productively without creating a negative classroom climate.

Keywords: race/ethnicity matching, teacher quality, culturally responsive teaching, experiment

Publication Note: A revised version of this working paper has been published in *Educational Researcher* under the title "Why Black Teachers Matter" (2024). Readers are encouraged to cite the final published version: Blazar, D. (2024). Why Black teachers matter. *Educational Researcher, 53*(8), 450–463. https://doi.org/10.3102/0013189X241261336

Correspondence regarding the paper can be sent to David Blazar at dblazar@umd.edu; Department of Teaching and Learning, Policy and Leadership, 2311 Benjamin Building, University of Maryland College of Education, College Park, MD, 20742. I thank staff in the research offices of collaborating districts that provided data; Casey Archer at the University of Maryland for research support; and Seth Gershenson, Ramon Goings, David Liebowitz, and Susanna Loeb for providing feedback on earlier versions of the paper.

Introduction

For far too long, education systems have failed students of color. Systemic racism—exhibited through school-based segregation (Johnson, 2011), exclusionary discipline (Fenning & Rose, 2007), limited access to instructional resources (Jackson, 2009), among other sources—has created stark disparities in educational opportunity between Black, Hispanic, and other historically marginalized and minoritized students of color versus their White peers. Constrained opportunity impacts and ripples across a range of educational and life outcomes, including academic performance (Fryer & Levitt, 2004; Jencks & Phillips, 2011; Reardon & Galindo, 2009), high school graduation (Hernandez, 2011; Murnane, 2013), college going (Merolla, 2018), and success in the labor market (Rivkin, 1995).

Compelling lines of theoretical and empirical research show that one of the most effective levers to better support students of color is to provide opportunities to learn from a teacher from the same racial or ethnic group. While the teacher workforce is overwhelmingly White (U.S. Department of Education, 2019), teachers of color are described as uniquely positioned to understand and address the social, political, and economic inequalities that students of color face (Irvine, 1989; Graham, 1987; Ladson-Billings, 1994; Waters, 1989). Building from this theory, causally oriented studies indicate that students of color who have a same-race/ethnicity are more likely to be held to high expectations for educational attainment (Gershenson et al., 2016), and are less likely to be perceived as disruptive or inattentive (Dee, 2005), be absent from school (Gottfried et al., 2021; Holt & Gershenson, 2019), or be suspended from school (Lindsay & Hart, 2017; Shirrell et al., 2021). In the only other experimental dataset—aside from the current study—used to examine the impact of teachers of color[1], Dee (2004) estimates effects of Black teachers on end-of-year test scores of Black students of more than 0.20

[1] Egalite and Kisida (2018) estimate race/ethnicity-matching effects in data from the Measures of Effective Teaching (MET) project, where a subsample of teachers was randomly assigned to class rosters within schools. However, the authors rely primarily on the observational/non-experimental portion of the dataset, noting that "noncompliance within the randomization procedure…tempers our ability to be absolutely certain that the effects we identify are causal" (p. 64).

standard deviations (SD), representing roughly one-third of the Black-White test score gap (Fryer & Levitt, 2004; Jencks & Phillips, 2011). In the same data, these short-term effects translate into key markers of human capital development for students of color, including increased rates of high school graduation and college going (4 to 5 percentage points; Gershenson et al., 2018).

The current analyses build from the prior literature by examining the underlying mechanisms driving average effects of teachers of color on student outcomes. To do so, I draw on a dataset where fourth- and fifth-grade teachers were randomly assigned to class rosters within schools, paired with rich data on varied student outcomes and varied teacher mindsets and practices. With these data, I first estimate the impacts of teachers of color on intrapersonal components of students' social-emotional development (i.e., self-efficacy, engagement, self-regulation), test scores in math and English language arts (ELA), and observed school behaviors (i.e., absences, suspensions), captured both at the end of the year working with that teacher and several years later when students are in high school. While theory suggests that the effects of teachers of color on students' academic performance likely works through intermediate effects on social-emotional development (Gay, 2000), empirical analysis of these relationships is limited. A handful of non-experimental studies shows positive associations between teacher-student race/ethnicity matching and student engagement in learning activities, motivation, and social skills (Egalite & Kisida, 2018; Rasheed et al., 2020; Wright et al., 2017). Building from this work, the experimental design in the current study allows for stronger causal claims. Further, linking the random assignment of teachers to classes in upper-elementary school to longer-run outcomes in high school provides one perspective on how short-term effects on social-emotional development and other outcomes translate into longer-run performance and experiences in school.

Next, I examine whether there are differences between teachers of color and their White colleagues in terms of mindsets and practices aligned to "culturally responsive teaching" (CRT; Gay, 2002), and whether these differences mediate average effects of teachers of color on student outcomes.

Throughout the paper, I refer to "CRT" but also recognize that the measures and dimensions in the data overlap with discussion of "culturally relevant pedagogy" (Ladson-Billings, 1995b) and "culturally sustaining pedagogy" (Paris, 2012). More specifically, the mindset and practice data—which come from teacher surveys and observations of instruction scored on an established observation instrument—capture the extent to which teachers hold growth mindset beliefs that student intelligence is malleable rather than fixed, develop strong interpersonal relationships with students and families, plan for and differentiate instruction for individual students' needs, and lead well-organized classrooms in which student (mis)behavior is addressed productively without creating a negative classroom climate. To date, the literature on CRT has focused primarily on qualitatively describing culturally responsive classroom contexts and practices, which suggests that students in these environments are more engaged than students in classrooms without these features (e.g., Ladson-Billings, 1995b; Milner, 2011; Ware, 2006). However, the research base aimed at causally linking CRT-aligned mindsets and practices to student outcomes is quite limited (Hill, 2020; Larson et al., 2018).

A secondary contribution of this paper is that it is only the second experimental study to examine the effect of teachers of color on student outcomes. This study focuses on random assignment of teachers to classes in fourth and fifth grade, whereas prior analyses using data from the Project STAR/Tennessee class size experiment focused on kindergarten through third grade (Dee, 2004; Gershenson et al., 2018). Replication of experimental effects is particularly useful given that estimated impacts of teachers of color on student test scores from the Project STAR data (roughly 0.2 SD) are substantially larger than estimates from most of the non-experimental studies (Redding, 2019).

Given the research intensity of randomly assigning teachers to classrooms and then collecting a broad set of teacher- and student-level measures beyond those typically captured in administrative datasets, the experimental sample size is moderate (n = 71 teachers and 1,283 students). At the same time, participants come from four school districts on the east coast of the U.S., and their characteristics

generally match those of the broader district populations. Sample size considerations also lead me to focus on teachers and students of color as a group (mostly Black, but also Hispanic and Asian; see Table 1). This approach aligns with recent, non-experimental studies showing benefits of teachers of color generally for students of color, rather than race/ethnicity-matching effects more narrowly (Blazar & Lagos, 2021; Lindsay & Hart, 2017; Shirrell et al., 2021).

In summary, I find that teachers of color have large and lasting effects on the social-emotional, academic, and behavioral outcomes of their students. Random assignment to a teacher of color in upper-elementary grades results in improved self-efficacy and classroom engagement (upwards of 0.45 SD), end-of-year math and ELA test scores (upwards of 0.26 SD), and school attendance (reductions in chronic absenteeism of 4 percentage points, representing a 60% decrease relative to students working with a White teacher). Short-term effects on test scores and chronic absenteeism persist at similar magnitudes up to six years later when students are in high school (roughly 0.2 SD on test scores, and 42% decrease in chronic absenteeism). Further, the effects of teachers of color on self-efficacy extend not just to students of color but also to their White peers, relative to White students working with a White teacher. This finding differs from other literature indicating that teachers of color generally have no impact—either positive or negative—on the test scores and school behaviors of White students, compared to having a White teacher (Gershenson et al., 2018; Shirrell et al., 2021).

These average effects of teachers of color on student outcomes are explained in part by specific mindsets and practices of teachers of color versus White teachers. Examining mean differences in teacher survey and classroom instructional quality measures, I find that teachers of color are more likely than their White colleagues to view student intelligence as malleable versus fixed, build interpersonal relationships with students and their families, spend more time planning for instruction and differentiating pedagogical approaches to individual students' needs, and lead well-organized classrooms. Differences are as large as 0.7 SD. In short, teachers of color are culturally responsive

teachers. Linking the teacher mindset and practice data to student outcomes, I find that teachers' growth mindset beliefs and their relationships with students and families serve as key mediators. Estimates from some models indicate that time planning for and differentiating instruction, as well as leading well-organized classrooms also serve as mediators. These patterns indicate that much of the effect of teachers of color on student outcomes runs through the intra- and interpersonal skills that teachers of color possess.

At the same time, in no model is the average effect of teachers of color on student outcomes fully explained by the available mediators. Given evidence of partial but not full mediation, it also is likely that other mechanisms play a role. As described in more detail below, CRT is a multidimensional enterprise (Gay, 2002; Ladson-Billings, 1994). Many but not all dimensions are measured in the data used in this study. Those dimensions that are more difficult to measure (e.g., development of critical consciousness) may also serve as mediators. Similarly, teachers of color likely serve as role models for students of color (Fordham & Ogbu, 1986). However, role-modeling effects are not as easily testable quantitatively, and generally rely on ruling out other possible mechanisms (Gershenson et al., 2018).

Motivating Literature

Although findings from the education production function literature (Monk, 1989; Todd & Wolpin, 2003) have crystallized around the benefit of same-race/ethnicity teachers for improving student outcomes (Bristol & Martin-Fernandez, 2019; Redding, 2019), less is known from this same research tradition about the mechanisms driving these effects. Theory, largely grounded in sociological and human development perspectives, suggest three possible pathways. First, students of color benefit from having teachers of color as role models, particularly given the way in which their career and training exemplifies academic success (Fordham & Ogbu, 1986). Increased diversity amongst schools' professional staff can help offset the normalization of race/ethnicity-based stratification that students of color experience in their lives both inside and outside of school (Villegas & Lucas, 2004).

Second—and not mutually exclusive from the first pathway—teachers of color may be better equipped than White teachers at teaching students of color. Scholars describe how the academic experiences and outcomes of students of color are informed by their lives beyond the classroom. As such, it is important that they have teachers who recognize and seek to understand how racial inequality shapes their world (Irvine, 1989; Ladson-Billings, 1994). White teachers are not inherently unable to teach students of color, but may be more likely than teachers of color to adopt and maintain deficit views and colorblind ideologies that presume that individual factors—rather than systemic racism—are responsible for the academic challenges that students of color may experience (Lewis, 2001; Valencia, 1997). Compared to White teachers, teachers of color may also have higher expectations for students of color (Ferguson, 2003), which can be critical for offsetting "stereotype threat" and the risk of confirming a negative stereotype about a group (Steele & Aronson, 1995).

Third, teachers of color may be better at teaching *all* students. Whether Black, Hispanic, Asian, or White, students report feeling better cared for and more academically challenged when they have a teacher of color (Cherng & Halpin, 2016). In other words, the practices and behaviors that teachers of color may deliver with students of color in mind may just be "good teaching" all around (Ladson-Billings, 1995a). There is a large literature, for example, describing how teachers' interpersonal relationships with students benefits an array of academic and social-emotional outcomes, on average across teachers and students from different backgrounds (Perlam et al., 2016; Pianta & Hamre, 2009).

However, scholars generally have been quite limited by available data to explore these mechanisms and mediating pathways in any rigorous way. The current study builds most directly from Gershenson et al.'s (2018) work, which exploits the random assignment of teachers to students in the Project STAR data to identify race/ethnicity-matching effects on short- and long-run student outcomes, as well as to explore possible mechanisms underlying these effects. The authors find that the effects of Black teachers on short-term test scores and longer-run outcomes at the end of high

school extend only to Black students and not to White students, which they interpret as evidence that Black teachers are not necessarily more effective overall. In further support of this claim, they find Black teacher-student matching effects are similar when including/excluding observable background characteristics of teachers (i.e., teaching experience, highest degree attained, status on a career ladder).

Yet, the authors also acknowledge that identifying exact mechanisms through which the effect of same-race/ethnicity teachers runs requires additional data. The background characteristics of teachers in their dataset are indirect proxies for effective teaching (Stronge, 2018). And, they do not align with the CRT literature that emphasizes five specific channels (Gay, 2000; Ladson-Billings, 1995b; Paris, 2012): (i) holding students to *high expectations* for academic learning and achievement; (ii) building strong *interpersonal relationships* with students to support engagement in the classroom environment; (iii) also building relationships with and understanding of students' lives outside of the classroom and then using this *cultural competency* to guide instruction; (iv) *differentiating instruction* to meet the needs of individual students; and (v) guiding students towards *critical consciousness* that allows them to critique cultural norms, values, and institutions that produce and maintain social inequities.

Over several decades, scholars have engaged in qualitative exploration of classrooms, providing rich descriptions of how teachers of color (and others) enact CRT, as well as their students' observed responses to these practices (e.g., Ladson-Billings, 1995b; Milner, 2011; Ware, 2006). More recently, research teams have developed observation instruments to identify and score the quality of teachers' classroom practice along different CRT-aligned dimensions (e.g., Goffney, 2010; Jensen et al., 2018; Powell et al., 2016). However, few studies have linked these observed classroom and teaching practices to student outcomes. One exception is a study by Larson et al. (2018), which finds a positive association between positive student behavior and a CRT measure that asked classroom observers to look for instances of strong teacher-student interactions, connections between content and real-world examples, integration of cultural artifacts into learning activities, and others. The authors note that,

like the broader literature base, their study is limited in its ability to draw causal conclusions because neither teachers nor the practices were randomly assigned to classes or students.

CRT includes not just teachers' classroom instruction but also their beliefs and expectations for students' academic performance and attainment (Gay, 2000; Ladson-Billings, 1995b; Paris, 2012). Aligned to this discussion, Papageorge et al. (2020) found racial biases in teachers' expectations for students' long-term educational attainment, with White teachers being substantially more optimistic about White students' ability to complete a four-year college degree compared to White teachers' beliefs on Black students' educational attainment. On average across teachers from different backgrounds, teachers' academic expectations for students affect actual college degree attainment. However, the authors do not examine directly whether differences in academic expectations between teachers of color and White teachers mediate effects on longer-run student outcomes.

There also is some relevant discussion of the effect of instructional programs (e.g., professional development, curricula) aimed at supporting teachers to deliver CRT, and subsequent effects on students. Here too, though, the evidence base is slim. In a recent comprehensive review, Bottiani et al. (2018) identified just two quantitative studies that compared outcomes of participants in a CRT-aligned intervention versus those not exposed to the intervention. One study focused on diversity training (Thompson & Byrnes, 2011), while the other focused on a schoolwide program that used data-based decision making and support from school leaders to help teachers create classroom environments with a focus on CRT (Vincent et al., 2011). However, both studies failed to meet evidence standards for supporting causal inferences. Since that review, Dee and Penner (2017) have found positive effects on student outcomes of an ethnic studies curriculum in California with features aligned to CRT, as well as an instructional and mentoring program for Black males developed through the Obama administration's My Brother's Keeper program (Dee & Penner, 2019). Both use causally-oriented research designs, but not randomized control trials.

Ultimately, scholars have called for more evidence—and more rigorous, causal evidence in particular—on the mechanisms linking teachers of color to improved student outcomes, and the extent to which CRT practices serve as key mediators (Bottiani et al., 2018; Gershenson et al., 2018; Hill, 2020). The current study aims to fill this gap.

Sample and Experimental Design

The data used in this study come from a research project called the National Center for Teacher Effectiveness, which examined characteristics of effective teachers and effective teaching in upper-elementary classrooms (i.e., fourth and fifth grade). A primary content focus of the study was mathematics, though participating teachers were generalists who taught all core subjects; data collection efforts and instruments crossed content areas. Over three school years (2010-11 through 2012-13), the research team collaborated with teachers (n = 321 total) in four school districts on the east coast of the U.S. to collect a variety of teacher-level measures through a set of teacher surveys and observations of classroom, and then to link these measures to researcher-collected student surveys and district-collected test scores, absences, and suspension records (see below for in-depth discussion of the teacher and student measures). The current analyses also link the project-organized data collection with administrative records on student test scores, absences, and suspensions provided by the partner districts through the 2018-19 school year (i.e., the last school year before the Covid-19 pandemic interrupted districts' collection of these measures).

During the third year of the study, a subset of teachers (n = 71) agreed to be randomly assigned to class rosters within schools. In spring 2012 the project team worked with staff at participating schools to randomly assign sets of teachers to class rosters (n = 1,283 students) of the same grade level that were constructed by principals or other school leaders. To be eligible for randomization teachers had to work in schools and grades in which there was at least one other participating teacher. Their principal also had to consider these teachers as capable of teaching any of the rosters of students

designated for the group of teachers.[2]

A moderately sized sample was the best way to capture breadth of information on students and teachers, as well as to ensure reasonably high compliance with the experimental design. Only a handful of other studies have randomly assigned teachers to classes in real-world school settings (e.g., Dee, 2004; Glazerman et al., 2006; Kane et al., 2013; Kane & Staiger, 2008). Of these, only the Measures of Effective Teaching (MET) project also collected a broad range of student- and teacher-level measures like those analyzed in this paper. However, in the MET study, the highest compliance rate of the six participating districts was 66% and the lowest rate was 27% (Kane et al., 2013). Egalite and Kisida (2018) explore teacher-student race/ethnicity-matching effects in these data but note that the high rate of noncompliance limits their ability to draw causal conclusions. In the current study, 69% of all students complied with the experimental design, and analyses presented below indicate that lingering noncompliance does not threaten the internal validity of results.

The project's volunteer sample matches the characteristics of teachers and students across the four participating school districts, and of urban school districts in the U.S. more broadly. In Table 1, I show that the subset of teachers in the experimental sample look similar to all upper-elementary teachers in their respective districts on their impacts on students' math test scores ($p = 0.687$). The subset of volunteer teachers for the experiment also look similar to the full project sample on a range of background teacher characteristics (i.e., gender, race/ethnicity, teaching experience), with no statistically significant differences. Teachers in the experiment were slightly more likely than those in the full research project to be certified through traditional rather than alternative programs. While

[2] I exclude four intact randomization blocks with 10 teachers who originally agreed to participate in the random assignment study but were missing relevant data for one of several reasons: four teachers left the study before the beginning of the 2012-13 school year for reasons unrelated to the experiment (i.e., leaving the district or teaching, maternity leave, change in teaching assignment); the principal of two teachers decided that it was not possible to randomly assign rosters to these teachers; and four teachers has random assignment partner(s) who left the study for either of the two reasons above. As randomization blocks are analogous to individual experiments, dropping individual ones does not threaten the internal validity of results.

these observable background teacher characteristics were not available for all teachers in the four partner districts, characteristics of the sample match national patterns (U.S. Department of Education, 2019). The vast majority of participating teachers were White (70%) and female (84%), with roughly 11 years of teaching experience. Of the participating teachers of color, the vast majority are Black (23% of the full experimental sample); 4% of teachers in the experiment are Asian, and 3% are Hispanic. I observe statistically significant differences across samples on several student characteristics. However, the magnitudes of these differences tend to be small.

In Table 2, I present estimates that confirm the success of the randomization process in creating balance between treatment groups. I find no difference between the characteristics of students whose randomly assigned teacher is White versus a teacher of color, both when relationships are tested individually and as a group ($p = 0.190$ on joint test of significance). In a related study using the same data, I also provide evidence that student characteristics are unrelated to the baseline effectiveness of students' randomly assigned teachers at raising math test scores (Blazar, 2018).

In a randomized experiment, the main threat to internal validity is attrition amongst participating students (Angrist & Pischke, 2009). In this study, attrition was driven by missing data, due to: (i) students moving out of their randomly assigned teachers' classroom, and so no longer were part of primary data collection; (ii) non-participation in district-led data collection; or (iii) moving out of the district or dropping out of high school, meaning that students are not observed in the longer-run high school data. As noted above, 31% of students moved out of their randomly assigned teacher's classroom. Importantly, though, these moves were unrelated to whether or not their assigned teacher was White versus a teacher of color ($p = 0.594$). Aligned to an intent-to-treat analysis (described below), I still include noncompliers in the analytic sample as long as they have outcome data.[3] In Table

[3] Of noncompliers, between 60% and 80% are missing outcome data: 64% and 72% of upper-elementary and high school test scores, respectively; 62% and 70% of upper-elementary and high school absence/suspension data, respectively; and

3, I show that there is no relationship between missingness—on each data source and time period—and whether or not their teacher is a person of color.

Data

The analyses presented in this paper draw on a combination of primary data collected as part of the research project and secondary data collected from districts' administrative records. I organize discussion around student- and teacher-level measures.

Students' Social-Emotional, Academic, and Behavioral Outcomes

District administrative data allow me to capture student demographic characteristics, end-of-year test scores, and observed school behaviors (e.g., absences, suspensions). These data were available during the research study (i.e., 2010-11 to 2012-13) and all subsequent years through 2018-19. Student demographic and background data include gender, race/ethnicity, free or reduced-price lunch (FRPL) eligibility, limited English proficiency (LEP) status, and receipt of special education (SPED) services.

Collection of test score data is driven by federal education policy that requires public school districts to administer end-of-year tests in math and ELA to all students in grades three through eight, and once in high school (see Lynch et al., 2017 for discussion of the upper-elementary content of each district/state test). In the experimental year (i.e., 2012-13), 78% of the fourth- or fifth-grade students had test scores (see Table 3); most of the students missing these tests moved out of the districts' public-school system after random assignment. Comparably, 52% of students in the experimental sample are linked to high school test scores. The drop in test-score coverage over time is due to additional students moving out of the district between upper-elementary and high school, as well as the fact that districts have discretion about the specific grade level in which they administer math and ELA exams in high school. Testing schedules also were interrupted by Covid starting in the 2019-20

79% of the project-administered student survey. Of all students missing outcome data, non-compliers account for 83% to 93% in upper-elementary school, and 47% to 52% in high school.

school year. As such, some students who still were enrolled in the same district in high school are missing longer-run test scores. I standardized all test scores within district, grade, and year to have a mean of 0 and a SD of 1. In Table 4a, I present univariate descriptive statistics on student outcome measures, showing that mean test-score performance of students in the experiment tend to be above district averages, and that there is less variation relative to the full district populations.

For observed school behaviors that also come from districts' administrative records, absences data tally the total number of days students were absent from school each year, and suspension data capture the number of days that students were suspended from school (either in-school suspension or out-of-school suspension). As long as students were enrolled in the school district in a given year, they have both absence and suspension records.[4] On average, upper-elementary students in the experiment missed 6.2 school days and were suspended a total of 0.09 days; in high school, students missed 10.2 days and were suspended 1 day, on average (see Table 4a). However, most students never were suspended, and many did not miss any days of school. Given the highly skewed nature of the absence and suspension data, I follow others in creating dichotomous measures that capture chronic absenteeism (i.e., missing 10% of total school days) and whether or not students were suspended at all in a given year (Gottfried, 2014; Holt & Gershenson, 2019; Lindsay & Hart, 2017; Jackson, 2018). In high school, I use chronic absenteeism and ever suspended measures captured in the most recent year/grade level available for each student. This is similar to the approach taken with test scores, where students differ in terms of the year/grade that they took required high school exams in math and ELA.

In addition to the measures collected by participating school districts, the data also include student-reported measures of intrapersonal components of social-emotional development that were collected on an end-of-year survey administered by the research project staff. Theory and exploratory

[4] Only those students who were suspended show up in districts' administrative files capturing in-school infractions. Aligned to district guidance, I assume that students who do not show up in the suspensions dataset but who were enrolled in the district that year were not suspended.

factor analyses (Blazar & Kraft, 2017) identify three constructs: (i) *Self-Efficacy* (10 items, internal consistency reliability [α] = 0.76) captures students' effort, initiative, and perception that they can complete tasks; (ii) *Engagement and Happiness in Class* (5 items, α = 0.82) asks students about their affect, happiness in, and enjoyment of class activities; and (iii) *Self-Regulation* (3 items, α = 0.74) captures the extent to which students regulate their behavior to align with teachers' expectations (see Appendix Table 1 for survey item text).[5] For each of these student-reported outcomes, I created final scales by reverse coding items with negative valence, averaging student responses across all available items within the construct, and then standardizing to mean of 0 and SD of 1 within the full project sample.[6]

In Table 4b, I show that the student-reported constructs of social-emotional development correlate with test-score performance, absences, and suspensions in expected directions. All three survey measures positively predict math and ELA test scores, with the highest correlation between *Self-Efficacy* and math test scores (ρ = 0.31). The student-reported measures also negatively correlate with absences and suspensions. Of the three student-reported measures, *Self-Regulation* correlates most strongly with suspensions (ρ = -0.18), which makes sense given that both aim to capture students' school and classroom (mis)behavior. Similarly, *Engagement and Happiness in Class* correlates more strongly with absences than with suspensions, which aligns with discussion of absenteeism as a proxy measure for students' engagement in school (Gottfried, 2014).

In Appendix Table 2, I show correlations between outcomes captured in upper-elementary versus high school, where patterns are similar to those shown in Table 4. For example, *Self-Efficacy, Engagement and Happiness in Class,* and *Self-Regulation* all positively correlate with high school test scores

[5] *Self-Efficacy* and *Engagement and Happiness in Class* are strongly correlated (ρ = 0.65; see Table 4b), and exploratory factor analyses suggest that these items may cluster together to form a single construct. However, review of item texts against the psychology literature from which they were derived provides theoretical justification for dividing them into two distinct constructs (see Blazar & Kraft, 2017). Sufficiently high reliability for each construct also supports this decision.

[6] Composite scores that average across raw responses are correlated at 0.99 and above with scales that incorporate weights from the factor analysis.

(ρ = 0.15 to 0.27) and negatively correlate with counts of high school absences (ρ = -0.07 to -0.12) and days suspended (ρ = -0.09 to -0.16). These patterns align with similar evidence showing the predictive validity of student-reported social-emotional development measures to long-run outcomes in college and the labor market (Lyubomirsky et al., 2005; Mueller & Plug, 2006).

Teacher Mindsets and Practices

Data on teachers come from a survey administered each fall during the research project, as well as videotaped observations of classes collected throughout each school year that subsequently were scored using the Classroom Assessment Scoring System (CLASS), an observation instrument shown to produce valid inferences regarding the quality of teachers' instruction (Bell et al., 2012). Given the focus of the research study on mathematics teaching, several survey items focus specifically on this content area, and all observations of classes come from math lessons. At the same time, as noted above, all participating teachers were generalists who taught all core subjects. Further, while the survey and CLASS measures were not developed specifically around CRT, several align closely with constructs described in the theoretical literature on this topic. In this paper, I focus on project-collected measures aligned to CRT, excluding others collected as part of the research study but not directly discussed in the CRT literature such as teachers' engagement in test-preparation activities.

More specifically, from the teacher survey I focus on three teacher-reported mindset and practice constructs. First is teachers' *Growth Mindset Beliefs* (7 items, α = 0.82), which captures the extent to which teachers view student intelligence as malleable versus fixed (Dweck, 2006).[7] The former perspective is thought to be most beneficial for students because it can support students' own

[7] Of the seven items from the *Growth Mindset Beliefs* construct, three were collected in one school year and four were collected in a different school year. Items were updated over time in light of initial psychometric analyses (α = 0.56 for the first set of items, and 0.93 for the second set of items). I calculate internal consistency reliability across all items by considering responses for teachers who completed the survey in both school years, and collapsing data to the teacher level. I include all seven items given the high pooled reliability estimate, and because use of all items leads to higher coverage across teachers in the sample. Patterns of results are very similar when excluding items from one year with lower reliability.

development of a growth mindset (Mesier et al., 2021) that, in turn, supports longer-run outcomes and academic success (Yeager & Dweck, 2020). Thus, I construct the *Growth Mindset Beliefs* measure such that higher scores reflect a mindset that student intelligence is more malleable. While this construct stems most directly from the psychology literature, there also are direct connections with the CRT literature, which describes how teachers should hold and then act on beliefs that "knowledge is not static" and that all students are capable of academic success no matter their prior background, experiences, and circumstances (Ladson-Billings, 1995b, p. 481). (See Table 5 for descriptive statistics, and Appendix Table 3 for survey item text.)

Second is teachers' *Relationships with Students and Families* (4 items, $\alpha = 0.63$), including the rapport teachers develop with students in and outside of the classroom, and the amount of time teachers spend talking with parents and families about students' learning and behavior.[8] For example, one item asks teachers the extent to which "students and I show an interest in each other's lives." This and the other three items align closely with discussion of CRT, which is grounded in strong social relations and instruction developed around students' lived experiences and backgrounds. The third construct from the teacher survey, *Preparation for Instruction* (14 items, $\alpha = 0.78$), identifies the amount of time teachers spend planning for instruction and collecting formative assessment data, as well as the extent to which they use this information to deliver differentiated instruction that attends to individual students' needs.[9] Scaffolding instruction is described in the early literature on CRT (Ladson-Billings, 1995a). More recently, scholars discuss how approaches to differentiation and data-driven

[8] One item from the *Relationships with Students and Families* construct—focused on teachers' interactions with family members—uses a different survey stem and response scale than the other items (see Appendix Table 3). Unsurprisingly, reliability is higher when excluding this item ($\alpha = 0.72$). I include the item given guidance from the CRT literature. Aligned to this discussion, the predictive power to student outcomes generally is stronger when including the additional item.

[9] The *Preparation for Instruction* items were developed to capture two constructs: out-of-class preparation and formative assessment. As shown in Appendix Table 3, each had its own response scale. However, exploratory factor analyses suggest that items cluster together to form a single construct. Theory on CRT also describes both types of practices as jointly facilitating teachers' knowledge of students and then delivery of student-oriented classroom instruction (Kieran & Anderson, 2019). From a practical perspective, moderate to strong correlation between the two constructs ($\rho = 0.48$) creates issues of multicollinearity when both are used as independent variables to predict student outcomes.

instruction have developed in tandem with CRT, all of which aim to better support historically marginalized and minoritized students (Kieran & Anderson, 2019; Santamaria, 2019). For all teacher-reported constructs, I averaged response across all available years of data, and standardized measures to have a mean of 0 and standard deviation of 1.

In addition to completing surveys, teachers contributed an average of three videotaped lessons per year[10], which trained raters scored on the CLASS instrument (Pianta et al., 2012). Exploratory and confirmatory factor analyses of CLASS scores in the data used here identify two constructs (Blazar et al., 2017). *Classroom Support* (9 items, $\alpha = 0.90$) focuses on teachers' interpersonal relationships with students around classroom activities and content, including creating a positive classroom climate, and teachers' sensitivity to and respect for student ideas and perspectives.[11] *Classroom Organization* (3 items, $\alpha = 0.72$) captures teachers' behavior management skills and the extent to which teachers' approach to addressing student (mis)behaviors avoids creating a negative classroom culture. Teachers' ability to productively organize and respond to student (mis)behavior is described as particularly important to building CRT-oriented classrooms given histories of exclusionary discipline for students of color (Fenning & Rose, 2007), as well as misunderstandings of the behavior, physical movements, and language of minoritized students (Gay, 2002). (See Appendix Table 4 for item text.)

Following protocols outlined by CLASS developers, I calculated teacher-level scores for *Classroom Support* and *Classroom Organization* by averaging scores across each 15-minute segment in a given lesson and across items within the dimension. Then, to account for variation in the number of

[10] Capture occurred with a freestanding, three-camera, digital recording device and lasted roughly minutes. One camera focused on the front of the classroom, while two others focused on student tables. Two microphones—one attached to the recording device and another worn by the teacher—picked up classroom talk. Teachers were allowed to choose dates, but were directed to select typical lessons and exclude days when students were taking a test. Although it is possible that these lessons captured instructional practice that were unique from a teachers' general instruction, teachers did not have any incentive to select lessons strategically. Analyses from separate data indicate that teachers are ranked almost identically when they choose lessons versus when lessons are chosen for them (Ho & Kane, 2013).

[11] *Classroom Support* combines items from the *Emotional Support* (4 items) and *Instructional Support* (5 items) dimensions originally outlined by CLASS instrument developers. Confirmatory factor analyses of the data used in this study indicate that these items cluster together to form a single construct (Blazar et al., 2017).

lessons teachers contributed to the dataset, I calculated predicted, shrunken teacher-level scores.[12] Using the raw data, I also calculated adjusted intraclass correlations (ICCs)[13] that capture the amount of construct-relevant variation at the teacher level. Scores of 0.63 and 0.47 for *Classroom Organization* and *Classroom Support*, respectively, are very similar to other large-scale studies that use the CLASS and other observation instruments to score classroom instruction (Bell et al., 2012; Kane & Staiger, 2012) and, thus, provide evidence of score reliability in addition to internal consistency reliability. I standardized observation scores to have a mean of 0 and a SD of 1.

In Table 5, I show pairwise correlations between the teacher mindset and practice measures. Unsurprisingly, relationships are strongest between constructs captured from the same measurement tool: ρ = 0.2 to 0.24 between survey measures, and 0.5 between the two dimensions derived from classroom observation and scored on the CLASS. Cross-instrument correlations are strongest between teacher-reported *Relationships with Students and Families* and the two dimensions from the CLASS observation instrument (ρ = 0.22 for *Classroom Support* and 0.28 for *Classroom Organization*). These patterns align with theoretical underpinnings of these measures, which all aim to capture teacher-student relationships and interactions, including but not limited to those that occur during instruction. Neither *Growth Mindset Beliefs* nor *Preparation for Instruction* are correlated with the two dimensions from the CLASS, indicating that these mindsets and practices do not translate directly into observed measures of high-quality classroom teaching captured by this observation instrument.

There is some degree of missingness in the teacher mindset and practices measures, given that teachers participated in the research project for different numbers of years and survey items varied to

[12] To estimate these scores, I specified the following multilevel model:

$$OBSERVATION_{ljt} = \gamma_l + \varepsilon_{ljt}$$

The outcome is the score for lesson *l* from teacher *j*. Teacher random effects, γ_l, are the parameters of interest, capturing mean observation scores for each teacher, shrunk to the sample mean based on the number of lessons for that teacher.

[13] The intra-class correlation (ICC) calculates the amount of variance in scores attributable to the teacher. Following a generalizability study framework (Hill et al., 2012), this ICC is adjusted for the median number of lessons per teacher.

some extent across years (see Appendix Table 5). Ninety-four percent of teachers in the experimental sample have observations of instruction scored on the CLASS, and 99% have scores on *Relationships with Students and Families*. Comparatively, 86% of teachers have data on *Growth Mindset Beliefs* and on *Preparation for Instruction*, where survey items only were included in the first two years of the research project. However, missingness is unrelated to whether or not a teacher is White or a person of color, or to other background characteristics (i.e., gender, teaching experience, certification pathway).

Empirical Strategy

The experimental design allows for a straightforward approach to estimate average effects of random assignment to a teacher of color on students' social-emotional, academic, and behavioral outcomes. All analyses began with a standard model of skill production (Monk, 1989; Todd & Wolpin, 2003) for student i in school s, and grade g, taught by teacher j in year t:

$$OUTCOME_{isgj(t+n)} = \beta TeacherOfColor_{ij} + \lambda_s + \tau_g + \varepsilon_{isgjt} \quad (1)$$

I use $OUTCOME_{isgj(t+n)}$ interchangeably for each student outcome, which I specify as a function of whether or not a student's randomly assigned teacher was White versus a teacher of color, $TeacherOfColor_{ij}$. Here, β captures the coefficient of interest, which I interpret as an average treatment effect given the intent-to-treat analyses. That is, the treatment variable captures whether or not a students' randomly assigned teacher was White versus a teacher of color, whether or not students stayed in or moved out of that classroom. It was not possible to instrument the race/ethnicity of students' actual teacher with their randomly assigned teacher, given that teacher race/ethnicity was not available for the full district population, as well as the fact that students who moved out of their randomly assigned teachers' classroom often switched districts and so are not observed in the data at all. However, as noted above, noncompliance and resulting missing data are unrelated to whether or not students' randomly assigned teacher is White versus a teacher of color (see Table 3). In the primary results, I focus on the single $TeacherOfColor_{ij}$ independent variable to generate a parsimonious set of

main effects on student outcomes. Parsimony is particularly desirable for the mediation analyses that incorporate several additional independent variables. In supplementary analyses, I also examine subgroup and interaction effects for students of color versus White students.

Outcomes are captured at the end of the year working with students' randomly assigned teacher (i.e., $t+n$, where $n = 1$), as well as several years later when students are in high school (i.e., $4 <= n <= 6$, depending on the grade level students were in during the experimental year and the last year they are observed in longer-run district data). To match the block randomized design, I control for fixed effects for school, λ_s, and grade, τ_g.[14] Inclusion of school fixed effects also helps account for differences in attendance/suspension reporting and state testing that can differ across schools and districts. I examine the sensitivity of estimates to inclusion versus exclusion of controls for background student and class characteristics (i.e., student-level variables listed in Table 2, and these same variables averaged to the class level plus class size), as well as background teacher characteristics (i.e., gender, certification pathway, teaching experience). My primary models are those that include background teacher characteristics, as this is where I have the greatest statistical power. I calculate heteroskedasticity-robust standard errors clustered at the teacher level to account for the clustered randomized design, with students nested within teachers' classrooms.

To examine causal mediation of CRT-aligned mindsets and practices, I pair estimates from equation (1) with additional estimates from a system of two equations that are similar to path analysis:

$$CRT_j = \delta TeacherOfColor_j + \mu_j \quad (2)$$

[14] In the experimental design, randomization blocks are equivalent to school-grade combinations. I expand identifying variation—by controlling for fixed effects for school and grade, rather than school-grade—given that not all school-grades with a teacher of color also have a White teacher; comparatively, all but one school with a teacher of color also has a White teacher. Including school and grade fixed effects rather than school-grade fixed effects does not appear to violate assumptions of baseline balance (see Table 2) or differential attrition (see Table 3). Patterns of results are quite similar and lead to similar conclusions when I control for school-grade fixed effects rather than school and grade fixed effects. In a sensitivity analysis described below, I also show that assumptions hold and patterns of results are similar when I replace school fixed effects with observable school characteristics, further expanding identifying variation.

$$OUTCOME_{isgj(t+n)} = \varphi TeacherOfColor_{ijt} + \pi CRT_j + \lambda_s + \tau_g + \varepsilon_{isgjt} \quad (3)$$

Equation (2) estimates differences in CRT-aligned mindsets and practices between teachers of color versus their White colleagues, which can be thought of as estimates of students' increased exposure to CRT-aligned mindsets and practices when assigned to a teacher of color. In the causal pathway, in equation (3), increased exposure to these mindsets and practices leads to improved student outcomes, conditional on whether or not students' randomly assigned teacher is White or a teacher of color. In other words, inclusion of the teacher mindset and practice measures in equation (3) may explain part (or all) of the average effect of teachers of color on student outcomes, as denoted in equation (1).

Mediation is difficult to examine causally because mediators often are not randomly assigned. For example, I cannot examine how effects on students' social-emotional development mediate effects on test scores, as the social-emotional outcome measures were not experimentally allocated to students. In the current study, though, I am interested in teacher-level mediators, and the level of randomization was the teacher. One important point on interpretation: I can interpret estimates of the relationship between a given mediator and student outcomes as the effect of being randomly assigned to a teacher who holds that mindset or engages in that practice. This interpretation is slightly different from the effect of random assignment of the mindset or practice itself.

For a given CRT-aligned mindset or practice to be considered a mediator, three conditions must be met: (i) there must be a relationship between CRT_j and $TeacherOfColor_j$ in equation (2); (ii) CRT_j must be related to student outcomes in equation (3), above and beyond the effect of the $TeacherOfColor_j$ variable on student outcomes; and (iii) the relationship between $TeacherOfColor_{ij}$ and student outcomes must differ in equation (3) relative to the magnitude of the estimate in equation (1) that excludes a given CRT_j measure as a mediator (Imai et al., 2010). To most directly assess these conditions, in my primary specification I examine mediation separately for each CRT-aligned mindset or practice. Because the mindsets and practices are correlated with each other (see Table 5), inclusion of all

measures in the same model predicting student outcomes makes it more difficult to assess whether changes in estimates of $TeacherOfColor_j$ from equation (1) to equation (3) are driven by one variable versus another. That said, I also specify a model of equation (3) that includes the full set of teacher mindset and practices as independent variables.

One concern with these mediation analyses is that CRT-aligned mindsets and practices—and those that are captured through classroom observations in particular—may be a function of the specific set of students in the classroom (Campbell & Ronfeldt, 2018; Steinberg & Garret, 2016). In other words, the relationship between CRT_j and $OUTCOME_{isgj(t+n)}$ in equation (3) may be artificially inflated if the same set of students contribute to both measures, and so are observed on both the left- and right-hand side of the equation. To avoid this concern, in equation (3) I focus on mindset and practice measures captured in years prior to the experiment.[15]

Results

Average Effects of Teachers of Color

In Table 6, I present estimates of the effect of random assignment to a teacher of color versus a White teacher on end-of-year outcomes in upper-elementary school in Panel A, and on longer-term high school outcomes in Panel B. All estimates come from separate regression models, each of which controls for school and grade fixed effects, and background teacher characteristics. For student-reported survey measures and test scores, higher scores reflect stronger outcomes. Estimates can be interpreted as the SD increase in these student outcomes resulting from being randomly assigned to a teacher of color relative to having a White teacher. Columns for social-emotional outcome measures

[15] *Growth Mindset Beliefs* and *Preparation for Instruction* items only were included in the teacher survey in years prior to the experiment. For *Relationships with Students and Families* and the two dimensions of classroom instructional quality from the CLASS instrument, scores are available for all three years of the research study. However, six teachers joined the study in the experimental year only, and so do not have prior-year scores. When these measures are used to predict student outcomes, I imputed prior-year scores using current-year scores. I justify this approach given high year-to-career correlations of 0.82 for *Relationships with Students and Families,* 0.72 for *Classroom Support*, and 0.77 for *Classroom Organization*. For teachers missing both prior- and current-year scores on any measure, I impute to the mean of the full sample given evidence that missingness is unrelated to background teacher characteristics (see Appendix Table 5).

in high school are blank, as the surveys measures were not collected by partner school districts in these grades. For chronic absenteeism and ever suspended variables that are measured dichotomously, a value of one indicates worse outcomes. In these linear probability models, estimates can be interpreted as the percentage point difference in being chronically absent or ever suspended in a given year as a result of being randomly assigned to a teacher of color.[16]

I find that random assignment to a teacher of color has large and lasting effects on students' social-emotional, academic, and behavioral outcomes. I find the largest effects of teachers of color on student-reported *Self-Efficacy* (0.45 SD). Effects on student-reported *Engagement and Happiness in Class* (0.25 SD) and on short-term math (0.26 SD) and reading test scores (0.23 SD) also are large; and they are similar to those found in a prior experiment focusing on younger students (Dee 2004). The test-score effects persist at very similar magnitudes up to six years later when students are in high school (0.18 to 0.22 SD). For behavioral outcomes, I find that random assignment to a teacher of color results in an 8-percentage point decrease in the probability of being chronically absent in high school. Relative to a 19% chronic absenteeism rate in high school for students whose randomly assigned teacher was White, 8 percentage points represents a 42% decrease.

Given the experimental design, estimates can be interpreted causally so long as baseline assumptions of balance between treatment groups and no differential attrition are met (see Tables 2 and 3). In Appendix Tables 6 and 7, I provide further evidence in support of the strength of the research design and subsequent inferences by varying the control set. Point estimates are quite similar when I add controls for background student and class characteristics (i.e., those characteristics listed in Table 2; see Appendix Table 6) in addition to background teacher characteristics, as well as when I remove all of these controls from the model (see Appendix Table 7). In models that add student and class characteristics, the estimate for decreased chronic absenteeism at the end of the year working

[16] For empirical justification of linear probability models relative to probit or logit models, see Heckman and Snyder (1996).

with a teacher of color of 4 percentage points is now statistically significantly different from zero. However, the estimate for decreased chronic absenteeism in high school is slightly smaller than in the primary results and no longer statistically significant. In the model without controls for background student, class, or teacher characteristics, I find that random assignment to a teacher of color results in increased suspensions (7 percentage points) over the course of the school year, compared to a 3% suspension rate for students working with a White teacher. In other models with different controls, this relationship also is positive in magnitude but not statistically significantly different from zero.

In Table 7, I examine heterogeneous effects across students of color versus White students, which provides additional insight into patterns related to suspensions. Models include four groups of teacher-student matches: (i) teachers of color randomly assigned to students of color, (ii) teachers of color randomly assigned to White students, (iii) White teachers randomly assigned to White students, and (iv) White teachers randomly assigned to students of color (the left-out category). Here, I find evidence of differential suspension rates of the students of White teachers, where their White students are suspended less frequently than their students of color (3.5 percentage points, representing an 88% decrease relative to a 4% suspension rate for students of color working with a White teacher).

Consistent with findings described above, suspension rates are higher both for White students and students of color assigned to a teacher of color, relative to the students of White teachers. White students are suspended more frequently when their randomly assigned teacher is a teacher of color versus a White teacher ($p = 0.031$; see panel in table testing equivalence of coefficients). There also is some suggestive evidence of differential suspension rates between White students and students of teacher randomly assigned to a teacher of color, given the magnitude of coefficients. However, the difference is not statistically significantly different from zero ($p = 0.173$). Further, when teachers of color are linked to high school suspensions, point estimates are negative in all model specifications and for both students of color and White students. But, none of the estimates is statistically

significantly different from zero. I explore additional patterns related to behavior in the mediation analyses, which include a measure of teachers' observed classroom management skills.

The subgroup analyses further reveal that teachers of color improve the *Self-Efficacy* of White students, relative to White students who have a White teacher ($p = 0.08$). For other outcome measures, I do not find statistically significant differences for White students between having a teacher of color versus a White teacher. Substantively, though, the effects of teachers of color on student-reported *Engagement and Happiness in Class* are concentrated amongst students of color. I find that random assignment to a teacher of color improves this measure by 0.31 SD, relative to students of color whose assigned teacher is White. Comparatively, the analogous point estimate for White students randomly assigned to a teacher of color is negative. That said, this estimate has a relatively large standard error and so is not statistically significantly different from zero, nor statistically significantly different from the effect of teachers of color on the classroom engagement and happiness of students of color.

Mediating Effects of CRT-Oriented Mindsets and Practices

To examine mechanisms underlying the average effects of teachers of color on student outcomes, I begin by examining differences in CRT-aligned mindsets and practices between teachers of color versus White teachers. In Table 8, I estimate mean differences in the full project sample, where I have the greatest statistical power. In Appendix Table 8, I show estimates of differences for teachers in the experimental sample. In both tables, I use data from all available school years. Sample sizes vary across measures, given that some measures were captured in a subset of years of the research study and not all teachers participated in the study for all years. I also specify models with various sets of controls, including background teacher characteristics (i.e., gender, certification pathway, teaching experience) and school fixed effects, which match the setup of models that predict student outcomes.

I find that teachers of color score higher on all five CRT-aligned measures relative to White teachers. In the full project sample, between-group differences are largest for *Preparation for Instruction*

(roughly 0.6 to 0.65 SD), and are very similar across models with different sets of controls. I also observe differences of 0.2 to 0.3 SD for *Growth Mindset Beliefs*, *Relationships with Students and Families*, and *Classroom Organization*. Here, point estimates and levels of statistical significance differ to some extent across models with different sets of controls. At the same time, patterns all point to teachers of color outperforming White teachers on these metrics, as expected given theory on CRT and purposeful selection of CRT-aligned mindset and practices measures. For *Classroom Support*, point estimates generally are positive but smaller than for other measures, and none are statistically significantly different from zero. While this measure has a lower adjusted intraclass correlation than *Classroom Organization*, lack of between-group differences for *Classroom Support* does not appear to be driven by measurement error; standard errors are quite similar to those for other measures.

In the experimental sample, between-group differences follow similar patterns compared to the full project sample (see Appendix Table 8). As expected, standard errors are roughly double in magnitude given smaller sample sizes. Here, the difference between teachers of color and White teachers are larger for *Relationships with Students and Families* (upwards of 0.73 SD), *Growth Mindset Beliefs* (upwards of 0.55 SD), and *Classroom Organization* (upwards of 0.55 SD). Estimates tend to be smaller in models that include school fixed effects, and none of the estimates in these models are statistically significantly different from zero. One explanation may be that variation in the measures is substantially reduced in the experimental sample and in models that include school fixed effects, compared to in the full project sample where there were many more teachers per school.

Next, in Table 9, I add the CRT-aligned mindset and practice measures to models that predict student outcomes. I focus on models that include just one CRT-aligned measure at a time in order to examine differences in the magnitude of effects of teacher of colors when including versus excluding the mindset and practice measure (see Table 6). Differences in estimates across these models is a key criterion for a given measure to be considered a mediator. In Appendix Table 9, I include all mindset

and practice measures as predictors in the same models, finding similar patterns. I exclude *Classroom Support* from these analyses, as I find no evidence of statistically significant differences in this measure between teachers of color and White teachers, either in the full project sample (see Table 8) or in the experimental sample (see Appendix Table 8). Between-group differences also is a key criterion for the measure to be considered a mediator of the effect of teachers of color on student outcomes. All teacher mindset and practice measures come from years prior to the experiment, and scores are imputed to the sample mean when missing so that all teachers and their students remain in the analysis. In Appendix Table 5, I show that missingness is unrelated to background teacher characteristics.

I find that *Growth Mindset Beliefs* and *Relationships with Students and Families* serve as mediators of the effect of teachers of color on varied student outcomes. Conditional on whether or not a students' randomly assigned teacher is White or a teacher of color, random assignment to a teacher who reports stronger relationships with students and families (i.e., 1 SD increase in this measure) results in improved *Self-Efficacy* (0.06 SD), improved longer-run math test scores in high school (0.08 SD), and decreased chronic absenteeism both in the short- and longer-term (2 percentage points). Assignment to a teacher who more strongly believes that student intelligence is malleable versus fixed results in decreased suspension rates at the end of the school year (2 percentage points) and several years later when students are in high school (3 percentage points), as well as decreased chronic absenteeism in high school (2 percentage points). In all of these models, estimates of the effect of having a teacher of color on student outcomes attenuate (i.e., are closer to zero) compared to models that exclude the mediators. Yet, in no model is attenuation of the main effect of teachers of color on student outcomes more than 50%.

I also find some evidence that *Preparation for Instruction* and *Classroom Organization* may serve as mediators. In Table 9, I find a marginally statistically significant relationship between *Preparation for Instruction* and decreased longer-term student absenteeism (4 percentage points); the relationship to

short-term student-reported *Engagement and Happiness in Class* also is substantively significant (0.14 SD) but not statistically significant. Similarly, for *Classroom Organization*, the relationship to short- and longer-term math achievement is substantively but not statistically significant (0.08 and 0.1 SD, respectively). When conditioning on the other mindset and practice measures (see Appendix Table 9), the estimate between *Classroom Organization* and short-term math test scores is marginally statistically significantly different from zero (0.09 SD). In these conditional analyses, I also find some evidence that *Classroom Organization* mediates effects on end-of-year suspensions. Here, I observe an *increase* in suspensions (2 percentage points), which is consistent with discussion above that all students of teachers of color are suspend more frequently than the students of White teachers. Because teachers of color score higher on *Classroom Organization* than White teachers, on average, the mediation analyses suggest that suspensions may be used to address classroom misbehavior in a way that does not detract from the classroom climate.

As noted above, it is possible that limiting variation in the teacher mindset and practice measures within schools is too restrictive (see Appendix Table 8). Therefore, in Appendix Table 10, I also examine mediating effects in models that replace school fixed effects with observable school characteristics (i.e., background student-level characteristics averaged to the school level). Student characteristics remain balanced between teachers of color and White teachers when replacing school fixed effects with observable school characteristics (p = 0.301 on test that background student characteristics jointly predict random assignment to a teacher of color versus a White teacher; analogous to the baseline balance test shown in Table 2). Here, I find very similar patterns linking *Growth Mindset Beliefs* and *Relationships with Students and Families* to student outcomes as in the primary specification. I also find marginally statistically significant relationships between *Preparation for Instruction* and students' math achievement (0.1 SD) and *Engagement and Happiness in Class* (0.13 SD) in the short term, as well as between *Classroom Organization* and end-of-year math achievement (0.1 SD).

In these models, random assignment to a teacher of color who scores 1 SD higher on *Relationships with Students and Families* further results in increased *Engagement and Happiness in Class* (0.06 SD).

Discussion

Pairing random assignment of teachers to classes with rich data on varied student outcomes and varied teacher mindsets and practices, this study extends prior quantitative literature identifying positive effects of teacher-student race/ethnicity matching, as well as primarily qualitative and theoretical literature defining core domains of CRT thought to be primary mechanisms by which teachers of color benefit students of color.

First and foremost, findings from this paper contribute the first causal evidence that teacher mindsets and practices aligned to CRT serve as mediators of the effect of teachers of color on intrapersonal components of students' social-emotional development, academic performance, and school behaviors. I find that teachers of color are more likely than White teachers—often substantially so—to hold growth mindset beliefs that student intelligence is malleable rather than fixed, build interpersonal relationships with students and families, prepare for instruction and differentiate activities based on students' individual needs, and address student (mis)behavior in productive ways that do not lead to a negative classroom climate. I also find evidence that all of these mindsets and practices further link to varied student outcomes. Given longstanding discussion of similar patterns in qualitative data (e.g., Ladson-Billings, 1995b; Milner, 2011; Ware, 2006), these findings may seem intuitive. Yet, exploration in experimental data is critical because it provides a direct way to test theories regarding whether and how CRT-aligned mindsets and practices benefit student outcomes.

At the same time that the evidence points to some specific dimensions of CRT that benefit student outcomes, I cannot rule out the possible effects of other dimensions not examined in this study. CRT is multifaceted, drawing on multiple experiences, knowledge bases, and dispositions of teachers, only some of which are measured quantitatively in the data used for this paper. For example,

the data do not include a measure of teaching towards critical conscientious, nor a student-level measure of this construct. Relatedly, the role of specific CRT-aligned mindsets and practices in generating productive and supportive classroom environments is not mutually exclusive with the benefits of having teachers of color as role models. Seeing role models who look like students of color in positions of power is described as beneficial, whether or not teachers engage in markedly different instructional practices (Fordham & Ogbu, 1986; Villegas & Lucas, 2004). The mediation analyses indicate that the available mindsets and practices of teachers partially but do not fully mediate the effect of teachers of color on student outcomes. In other words, in models that include teacher mindset and practices, the effect of a teacher of color on student outcomes is attenuated relative to models that exclude these measures; but, the estimated effects are not zero. Therefore, it is very likely that other mechanisms such as role modeling and developing critical consciousness also play a role.

Further aligned to the theoretical literature (Gay, 2000), I find that teachers of color have the largest effects on the intrapersonal components of social-emotional development of students of color, including measures of self-efficacy (upwards of 0.47 SD) and engagement in class activities (upwards of 0.31 SD). Teachers of color also have large positive effects on test scores, but at smaller magnitudes (upwards of 0.26 SD). (Teachers of control also impact observed school behaviors, though these impact estimates are not directly comparable because the effects are measured in percentage points rather than in SD units.) These patterns provide suggestive evidence that the effects of teachers of color on test scores may partially run through more proximal effects on social-emotional development. However, this conclusion is not definitive given the experimental design: I can estimate the effect of random assignment to a teacher of color on both social-emotional development and academic performance, but I cannot estimate the effect of one of the outcomes on the other.

Focusing more narrowly on the effects of teachers of color on student test scores, there are two additional takeaways from this study. First, I find very similar end-of-year test-score effects of

teachers of color as Dee (2004) and replicated by Gershenson et al. (2018) who also draw on experimental data. In comparison, a growing number of non-experimental studies also find positive test-score effects of teacher-student race/ethnicity-matching, but at magnitudes that generally are much smaller (for reviews, see Bristol & Martin-Fernandez, 2019; Redding, 2019). One explanation may be related to generalizability. For example, Egalite et al. (2015) use statewide data from Florida and find test-score effects of 0.02 SD for Black students matched with a Black teacher. Analyses of statewide data have stronger generalizability than experiments with volunteer samples. At the same time, I show that the volunteer sample in this study looks similar to broader populations of four east coast school districts. The other explanation for differences across experimental versus non-experimental studies is the strength of the research design. In the absence of experimental data, researchers often include a combination a school and student fixed effects, which aim to account for non-random sorting of students to teachers that are well documented in schools (Clotfelter et al., 2006). Descriptive analyses tend to show teachers of color assigned to classes where students have below-average test-score performance and above-average absence and suspension rates (Holt & Gershenson, 2019; Kalogrides et al., 2013), meaning that lingering biases in student and school fixed effects models may understate true effects of teachers of color on student outcomes. Preferencing the experimental estimates, assignment to a teacher of color produces some of the largest effects on the outcomes of students of color across all of the education intervention literature (Fryer, 2017).

Second, whereas the effect of teachers and of educational interventions on students' academic performance tend to fade out substantially over time (Cascio & Staiger, 2012), I find that the effects of teachers of color on end-of-year tests scores in math and ELA persist at very similar magnitudes up to six years later when students are in high school. These test-score effects should not be interpreted as the main finding of this paper. Test scores capture just a small slice of the knowledge bases, skills, and dispositions students need when they leave school (Deming, 2017; Jackson, 2018),

and improving test scores generally is not the main goal of teacher-student race/ethnicity matching discussed in the theoretical literature (Irvine, 1989; Graham, 1987; Ladson-Billings, 1994; Waters, 1989). Yet, the degree of persistence of the test-score effects is striking relative to other analyses in the educational effectiveness literature.

This study also contributes to the literature on racial discrimination. Descriptively, nationally representative survey data show how Black teachers exhibit substantially less implicit racial bias than White teachers (Chin et al., 2020). Lab experiments that randomly assign student race to classroom assignments also show that White teachers grade Black students more harshly than White students (Quinn, 2020), which is similar to findings of discrimination in hiring practices (Quillian et al., 2017). In real-world schools and classrooms, racial biases are instantiated in lower teacher expectations for students' educational attainment (Papageorge et al., 2020), decreased access to advanced coursework (Tyson et al., 2007), and higher rates of suspensions and exclusionary discipline (Fenning & Rose, 2007), all of which limit academic opportunities and successes for students of color. This study extends the extant literature by providing experimental estimates of differential suspension rates by teacher and student race/ethnicity. Consistent with other literature, I find that random assignment to a White teacher results in higher suspension rates for students of color relative to their White peers. Though not statistically significantly different from zero, I also observe substantively higher suspension rates of White students relative to students of color when their randomly assigned teacher is a person of color. While school leaders ultimately make the final suspension decision, referrals most frequently come from teaches regarding incidences in their classrooms (Girvan et al., 2017; Liu et al., 2021). At the same time, teachers of color score substantially higher than White teachers in terms of their classroom management skills, suggesting that possible differential suspension rates of the students of teachers of color may not reflect exclusionary practices that lead to a poor classroom climate.

Conclusion

Ultimately, this paper aims to provide new knowledge and direction to the intervention, practice, and policy literatures regarding strategies for offsetting racial/ethnic discrimination in school and increasing opportunities for Black, Hispanic, and other historically marginalized and minoritized students of color. As noted above but worth repeating, the effects of teachers of color on students' social-emotional development, academic performance, and school behaviors are some of the largest across all of the experimental literature in education. As a general point of comparison, Fryer (2017) conducted a meta-analysis of experimental effects of a broad range of human-capital oriented interventions on student test scores, finding the largest effects for one-on-one, high-dosage tutoring (roughly 0.2 to 0.4 SD). The effects of teachers of color in this study are very similar in magnitude, and extend to a range of students in the classroom, rather than being a one-on-one intervention like tutoring. However, at present, the re-assignment of students of color to teachers of color is not a policy or practice intervention in and of itself, largely because teachers of color are wildly underrepresented in the teacher workforce (U.S. Department of Education, 2019).

How then can schools and districts, policymakers and practitioners act on these findings? Gershenson et al. (2018) describe how policy action depends largely on the underlying mechanisms by which teachers of color and race/ethnicity matching benefit students. If the effects are driven primarily or exclusively by role modeling, then the only real option is to engage in different approaches to recruitment and retention of teachers of color to substantially alter the demographics of the population of public-school teachers. Alternatively, training the mostly-White teacher workforce to deliver CRT may be an option if the effects of teachers of color are explained by the unique set of skills that they bring to their work. I find empirical support for the role that CRT-aligned mindsets and practices play in mediating the effects of teachers of color on student outcomes. As such, there may be some or even substantial benefit of allocating professional development dollars towards

shifting the mindsets of White teachers to a growth-mindset perspective and to help them engage in CRT-aligned practices. This suggestion also aligns with recent evidence on the positive effects of an ethnic studies curriculum aligned to CRT and that was implemented by teachers of all races and ethnicities in California (Bonilla et al., 2021; Dee & Penner, 2017).

However, in light of racial biases instantiated in differential suspension rates, I cannot rule out the possibility that some of the skills that teachers of color have, on average, and that benefit student outcomes may not be teachable—or, at least, not easily taught. Thus, recruitment of a more diverse teacher workforce must also be a key avenue for schools and districts. To date, we know hardly anything about the causal effect of policies and programs aimed at recruiting more individuals of color into teaching.

Stating that we need to diversify the teacher workforce is neither new nor novel. The evidence presented in this paper certainly points in that direction, while also providing potential avenues for policy and practice experimentation, and for partnership-based research. Are the CRT-oriented mindsets and practices examined in this study that mediate the effect of teachers of color on student outcomes teachable to White teachers? If so, how should these learning activities be designed? What are the programs, support systems, and incentives that can increase the share of individuals of color entering the teaching profession? These questions were first posed over 40 years ago. It is time that we find the answers.

References

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics*. Princeton University Press.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*, *13*(3), 281-309.

Blazar, D., Braslow, D., Charalambous, C.Y. and Hill, H.C., 2017. Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, *22*(2): 71-94.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146-170.

Blazar D., & Lagos, F. (2021). Professional staff diversity and student outcomes: Extending our understanding of race/ethnicity-matching effects in education. *EdWorking Paper No. 21-500*. Retrieved from: https://edworkingpapers.com/ai21-500

Bonilla, S., Dee, T. S., & Penner, E. K. (2021). Ethnic studies increases longer-run academic engagement and attainment. *Proceedings of the National Academy of Sciences*, *118*(37).

Bottiani, J. H., Larson, K. E., Debnam, K. J., Bischoff, C. M., & Bradshaw, C. P. (2018). Promoting educators' use of culturally responsive practices: A systematic review of inservice interventions. *Journal of Teacher Education, 69*(4), 367-385.

Bristol, T. J., & Martin-Fernandez, J. (2019). The added value of Latinx and Black teachers for Latinx and Black students: Implications for policy. *Policy Insights from the Behavioral and Brain Sciences*, *6*(2), 147-153.

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

Cherng, H. Y. S., & Halpin, P. F. (2016). The importance of minority teachers: Student perceptions of minority versus White teachers. *Educational Researcher*, *45*(7), 407-420.

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., & Lovison, V. S. (2020). Bias in the air: A nationwide exploration of teachers' implicit racial attitudes, aggregate bias, and student outcomes. *Educational Researcher*, *49*(8), 566-578.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of human Resources*, *41*(4), 778-820.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*(1), 195-210.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter?. *American Economic Review*, *95*(2), 158-165.

Dee, T. S., & Penner, E. K. (2017). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal, 54*(1), 127-166.

Dee, T., & Penner, E. (2019). *My brother's keeper? The impact of targeted educational supports* (No. w26386). National Bureau of Economic Research.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, *132*(4), 1593-1640.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House Digital, Inc.

Eddy, C. M., & Easton-Brooks, D. (2011). Ethnic matching, school placement, and mathematics achievement of African American students from kindergarten through fifth grade. *Urban Education*, *46*(6), 1280-1299.

Egalite, A. J., & Kisida, B. (2018). The effects of teacher match on students' academic perceptions and attitudes. *Educational Evaluation and Policy Analysis*, *40*(1), 59-81.

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, *45*, 44-52.

Fenning, P., & Rose, J. (2007). Overrepresentation of African American students in exclusionary discipline the role of school policy. *Urban Education, 42*(6), 536-559.

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the Black-White test score gap. *Urban Education*, *38*(4), 460-507.

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the "burden of 'acting white"'. *The Urban Review*, *18*(3), 176-206.

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments* (Vol. 2, pp. 95-322). North-Holland.

Fryer Jr, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, *86*(2), 447-464.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education, 53*(2), 106-116.

Gay, G. (2000). *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press.

Gershenson, S., Hart, C. M., Hyman, J., Lindsay, C., & Papageorge, N. W. (2018). *The long-run impacts of same-race teachers* (No. w25254). National Bureau of Economic Research.

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, *52*, 209-224.

Girvan, E. J., Gion, C., McIntosh, K., & Smolkowski, K. (2017). The relative contribution of subjective office referrals to racial disproportionality in school discipline. *School Psychology Quarterly*, *32*(3), 392.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management*, *25*(1), 75-96.

Goffney, I. D. (2010). *Identifying, measuring, and defining equitable mathematics instruction* (Doctoral dissertation).

Gottfried, M., Kirksey, J. J., & Fletcher, T. L. (2021). Do High School Students With a Same-Race Teacher Attend Class More Often?. *Educational Evaluation and Policy Analysis*, 01623737211032241.

Gottfried, M. A. (2014). Chronic absenteeism and its effects on students' academic and socioemotional outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, *19*(2), 53-75.

Graham, P. A. (1987). Black teachers: A drastically scarce resource. *The Phi Delta Kappan*, *68*(8), 598-605.

Heckman, J. J., & Snyder, J. M. (1996). *Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators.* NBER Working Paper No. 5785.

Hernandez, D. J. (2011). *Double jeopardy: How third-grade reading skills and poverty influence high school graduation.* Annie E. Casey Foundation. Retrieved from: https://eric.ed.gov/?id=ED518818

Hill, H. C. (2020, March). *Culturally responsive teaching is promising. But there's a pressing need for more research.* Education Week. *https://www.edweek.org/leadership/opinion-culturally-responsive-teaching-is-promising-but-theres-a-pressing-need-for-more-research/2020/03*

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Ho, Andrew D., and Thomas J. Kane. (2013). The reliability of classroom observations by school personnel. *Bill & Melinda Gates Foundation*.

Holt, S. B., & Gershenson, S. (2019). The impact of demographic representation on absences and suspensions. *Policy Studies Journal*, *47*(4), 1069-1099.
Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309.
Irvine, J. J. (1989). Beyond role models: An examination of cultural influences on the pedagogical perspectives of Black teachers. *Peabody Journal of Education, 66*(4), 51-63.
Irvine, J. J., & Armento, B. J. (2001). Culturally responsive teaching: Lesson planning for elementary and middle grades. *Education Review*.
Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics*, *27*(2), 213-256.
Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072-2107.
Jencks, C., & Phillips, M. (1998). *The Black-White test score gap*. Brookings Inst Press.
Jensen, B., Grajeda, S., & Haertel, E. (2018). Measuring cultural dimensions of classroom interactions. *Educational Assessment, 23*(4), 250-276.
Johnson, Rucker C. *Long-run impacts of school desegregation & school quality on adult attainments*. No. w16664. National Bureau of Economic Research, 2011.
Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103-123.
Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.
Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.
Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.
Kieran, L., & Anderson, C. (2019). Connecting universal design for learning with culturally responsive teaching. *Education and Urban Society, 51*(9), 1202-1216.
Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1-36.
Ladson-Billings, G. (1994). Who will teach our children: Preparing teachers to successfully teach African American students. *Teaching Diverse Populations: Formulating a Knowledge Base*, 129-142.
Ladson-Billings, G. (1995a). But that's just good teaching! The case for culturally relevant pedagogy. *Theory into practice*, *34*(3), 159-165.
Ladson-Billings, G. (1995b). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465-491.
Larson, K. E., Pas, E. T., Bradshaw, C. P., Rosenberg, M. S., & Day-Vines, N. L. (2018). Examining how proactive management and culturally responsive teaching relate to student behavior: Implications for measurement and practice. *School Psychology Review, 47*(2), 153-166.
Lewis, A. E. (2001). There is no "race" in the schoolyard: Color-blind ideology in an (almost) all-white school. *American Educational Research Journal*, *38*(4), 781-811.
Lindsay, C. A., & Hart, C. M. (2017). Exposure to same-race teachers and student disciplinary outcomes for Black students in North Carolina. *Educational Evaluation and Policy Analysis*, *39*(3), 485-510.

Liu, L., Hayes, M.S., & Gershenson, S. (2021). From referrals to suspensions: New evidence on racial discrimination in exclusionary discipline. *EdWorkingPaper No. 21-442*. Retrieved from: https://www.edworkingpapers.com/sites/default/files/ai21-442.pdf

Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between observations of elementary mathematics instruction and student achievement: Exploring variability across districts. *American Journal of Education*, *123*(4), 615-646.

Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success?. *Psychological Bulletin*, *131*(6), 803.

Merolla, D. M. (2018). Completing the educational career: High school graduation, four-year college enrollment, and bachelor's degree completion among Black, Hispanic, and White students. *Sociology of Race and Ethnicity*, *4*(2), 281-297.

Mesler, R. M., Corbin, C. M., & Martin, B. H. (2021). Teacher mindset is associated with development of students' growth mindset. *Journal of Applied Developmental Psychology, 76*, 101299.

Milner, H. R. (2011). Culturally relevant pedagogy in a diverse urban classroom. *The Urban Review, 43*(1), 66-89.

Monk, D. H. (1989). The education production function: Its evolving role in policy analysis. *Educational Evaluation and Policy Analysis*, *11*(1), 31-45.

Mueller, G., & Plug, E. (2006). Estimating the effect of personality on male and female earnings. *ILR Review*, *60*(1), 3-22.

Murnane, R. J. (2013). US high school graduation rates: Patterns and explanations. *Journal of Economic Literature*, *51*(2), 370-422.

Papageorge, N. W., Gershenson, S., & Kang, K. M. (2020). Teacher expectations matter. *Review of Economics and Statistics*, *102*(2), 234-251.

Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational Researcher*, *41*(3), 93-97.

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the classroom assessment scoring system) in early childhood education and care settings and child outcomes. *PloS one, 11*(12), e0167660.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109-119.

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Classroom assessment scoring system: Upper elementary manual*. Teachstone.

Powell, R., Chambers Cantrell, S., Malo-Juvera, V., & Correll, P. (2016). Operationalizing culturally responsive instruction: Preliminary findings of CRIOP research. *Teachers College Record, 118*(1), 1-46.

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, *114*(41), 10870-10875.

Quinn, D. M. (2020). Experimental evidence on teachers' racial bias in student evaluation: The role of grading scales. *Educational Evaluation and Policy Analysis*, *42*(3), 375-392.

Rasheed, D. S., Brown, J. L., Doyle, S. L., & Jennings, P. A. (2020). The effect of teacher–child race/ethnicity matching and classroom diversity on children's socioemotional and academic skills. *Child Development*, *91*(3), e597-e618.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, *46*(3), 853-891.

Redding, C. (2019). A teacher like me: A review of the effect of student–teacher racial/ethnic matching on teacher perceptions of students and student academic and behavioral outcomes. *Review of Educational Research*, *89*(4), 499-535.

Rivkin, S. G. (1995). Black/white differences in schooling and employment. *Journal of Human Resources*, 826-852.

Santamaria, L. J. (2009). Culturally responsive differentiated instruction: Narrowing gaps between best pedagogical practices benefiting all learners. *Teachers College Record, 111*(1), 214-247.

Shirrell, M., Bristol, T. J., & Britton, T. A. The effects of student-teacher ethnoracial matching on exclusionary discipline for Asian American, Black, and Latinx Students: Evidence from New York City. *EdWorkingPaper No. 21-475*. Retrieved from: https://edworkingpapers.org/sites/default/files/ai21-475.pdf

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797.

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.

Stronge, J. H. (2018). *Qualities of effective teachers*. ASCD.

Thompson, J., & Byrnes, D. (2011). A more diverse circle of friends. *Multicultural Perspectives, 13*(2), 93-99.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3-F33.

Tyson, W., Lee, R., Borman, K. M., & Hanson, M. A. (2007). Science, technology, engineering, and mathematics (STEM) pathways: High school science and math coursework and postsecondary degree attainment. *Journal of Education for Students Placed at Risk*, *12*(3), 243-270.

U.S. Department of Education. (2019). *The condition of Education.*

Valencia, Richard. (1997). *The evolution of deficit thinking*. London, United Kingdom: Falmer.

Villegas, A. M., & Lucas, T. F. (2004). Diversifying the teacher workforce: A retrospective and prospective analysis. *Yearbook of the National Society for the Study of Education*, *103*(1), 70-104.

Vincent, C. G., Randall, C., Cartledge, G., Tobin, T. J., & Swain-Bradway, J. (2011). Toward a conceptual integration of cultural responsiveness and schoolwide positive behavior support. *Journal of Positive Behavior Interventions, 13*(4), 219-229.

Ware, F. (2006). Warm demander pedagogy: Culturally responsive teaching that supports a culture of achievement for African American students. *Urban Education, 41*(4), 427-456.

Waters, M. M. (1989, September). An agenda for educating Black teachers. In The *Educational Forum* (Vol. 53, No. 3, pp. 267-279). Taylor & Francis Group.

Wright, A., Gottfried, M. A., & Le, V. N. (2017). A kindergarten teacher like me: The role of student-teacher race in social-emotional development. *American Educational Research Journal*, *54*(1_suppl), 78S-101S.

Yeager, D. S., & Dweck, C. S. (2020). What can be learned from growth mindset controversies?. *American Psychologist, 75*(9), 1269.

Tables

Table 1
Demographic Characteristics of Teachers and Students

	Experimental Sample	Full Project Sample		Full District Population	
Teacher Characteristics					
Female	0.85	0.78			
Teacher of Color	0.30	0.33			
Asian	0.04	0.03			
Black	0.23	0.22			
Hispanic	0.03	0.03			
White	0.70	0.65			
Traditionally Certified	0.92	0.85	~		
Alternatively Certified	0.05	0.08			
No Formal Training	0.03	0.07			
Teaching Experience (years)	11.05	10.59			
Teacher Effects on State Math Test (standardized)	-0.004	0.016	***	0.001	
Teachers	71	321		3,559	
Student Characteristics					
Female	0.47	0.50	~	0.50	~
Student of Color	0.80	0.76	***	0.75	***
Asian	0.10	0.08	*	0.09	
Black	0.41	0.40		0.36	***
Hispanic	0.24	0.23		0.27	*
White	0.20	0.24	***	0.25	***
FRPL	0.68	0.64	**	0.63	***
SPED	0.07	0.11	***	0.14	***
LEP	0.18	0.20		0.19	
Math Achievement (standardized)	0.038	0.101	*	0.008	
ELA Achievement (standardized)	0.050	0.084		0.006	
Students	1,283	10,586		175,572	

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1 on differences between the experimental sample and the full project sample or the full district populations.

Table 2
Balance Between Students Randomly Assigned to a Teacher of Color versus a White Teacher

Student Characteristics	Average Student Characteristics for Teachers of Color	Difference for White Teachers
Female	0.49	-0.009
		(0.011)
Asian	0.14	-0.066
		(0.043)
Black	0.32	0.004
		(0.035)
Hispanic	0.27	-0.021
		(0.031)
White	0.22	-0.047
		(0.032)
FRPL	0.67	0.004
		(0.011)
SPED	0.06	0.023
		(0.038)
LEP	0.27	-0.025
		(0.026)
Prior Math Achievement (standardized)	0.116	-0.016
		(0.015)
Prior ELA Achievement (standardized)	0.014	0.018
		(0.017)
Prior Absences (days)	5.6	-0.000
		(0.002)
Prior Suspensions (days)	0.08	-0.021
		(0.017)
P-value on Joint Test of Significance		0.190
Teachers	71	
Students	1,283	

Notes: Average student characteristics of teachers of color are calculated from regression models that control for school and grade fixed effects. Differences in average student characteristics for White teachers are calculated from a single regression model that predicts a dummy variable for the teacher being White as function of the student characteristics and school and grade fixed effects. Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses.

Table 3
Availability of Student Outcome Data and Relationship to Random Assignment to a Teacher of Color versus a White Teacher

	Upper-Elem. School		High School	
	Proportion with Data	P-Value	Proportion with Data	P-Value
Project Survey	0.70	0.664	NA	NA
Test Score Data	0.78	0.313	0.52	0.361
Absence/Suspension Data	0.79	0.442	0.58	0.259
Students	1,283			

Notes: P-values test the hypothesis that attrition/missingness is related to whether or not a student was randomly assigned to a teacher of color versus a White teacher, and are calculated from models that control for school and grade fixed effects, and cluster standard errors at the teacher level.

Table 4a
Descriptive Statistics for Student Outcomes

	Reliability	Upper-Elem. School		High School	
		Mean	SD	Mean	SD
<u>Intrapersonal Competencies</u>					
Self-Efficacy (1 to 5 scale)	0.76	4.09	0.64	NA	
Engagement/Happiness in Class (1 to 5 scale)	0.82	3.99	0.90	NA	
Self-Regulation (1 to 5 scale)	0.74	4.09	0.93	NA	
<u>Test Scores</u>					
Math Achievement (standardized)	>0.90	0.04	0.92	0.06	0.92
ELA Achievement (standardized)	>0.90	0.05	0.90	0.19	0.85
<u>School Behaviors</u>					
Absences (days)	NA	6.19	6.87	10.15	15.23
Chronically Absent	NA	0.05	NA	0.17	NA
Suspensions (days)	NA	0.09	0.54	0.95	3.06
Ever Suspended	NA	0.04	NA	0.18	NA

Notes: Reliability statistics for survey measures are internal consistency reliabilities. Reliability statistics for math and ELA tests vary across participating school districts and grade levels, but all are above 0.9.

Table 4b

Pairwise Correlations between Upper-Elementary Outcomes

	Self-Efficacy	Engagement Happiness in Class	Self-Regulation	Math Achievement	ELA Achievement	Absences	Chronically Absent	Suspensions	Ever Suspended
Intrapersonal Competencies									
Self-Efficacy	1								
Engagement/Happiness in Class	0.647***	1							
Self-Regulation	0.321***	0.241***	1						
Test Scores									
Math Achievement	0.312***	0.216***	0.217***	1					
ELA Achievement	0.245***	0.146***	0.275***	0.662***	1				
School Behaviors									
Absences (days)	-0.109**	-0.102**	-0.016	-0.183***	-0.094**	1			
Chronically Absent	-0.072*	-0.088**	-0.011	-0.137***	-0.060~	0.695***	1		
Suspensions	-0.108**	-0.091**	-0.151***	-0.086**	-0.062*	0.104**	0.043	1	
Ever Suspended	-0.087**	-0.072*	-0.177***	-0.112**	-0.111*	0.113**	0.054~	0.743***	1

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Table 5

Descriptive Statistics for Upper-Elementary Teacher Practices and Mindsets

	Reliability	Mean	SD	Pairwise Correlations				
				GMB	RSF	PFI	CS	CO
Growth Mindset Beliefs (GMB; 1 to 6 scale)	0.82	4.39	0.94	1.00				
Relationships with Students and Families (RSF; 1 to 5 scale)	0.63	3.86	0.63	0.196***	1			
Preparation for Instruction (PFI; 1 to 5 scale)	0.78	3.27	0.48	0.215**	0.236***	1		
Classroom Support (CS; 1 to 7 scale)	0.90; 0.47	4.12	0.42	0.051	0.222**	0.082	1.000	
Classroom Organization (CO; 1 to 7 scale)	0.72; 0.63	6.40	0.41	0.029	0.280***	0.003	0.514***	1

Notes: All measures have internal consistency reliability statistics; classroom observation scores also have adjusted intra-class correlations. *** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Table 6

Effect of Teachers of Color on Short- and Longer-Term Student Outcomes

	Self-Efficacy	Engagement/ Happiness in Class	Self-Regulation	Math Achievement	Reading Achievement	Chronically Absent	Ever Suspended
Panel A: End-of-Year Effects in Upper-Elementary School							
Tch. Of Color	0.448***	0.251~	0.123	0.257**	0.231*	-0.040	0.032
	(0.115)	(0.141)	(0.115)	(0.084)	(0.091)	(0.025)	(0.027)
Students	903	903	901	1,003	1,004	1,017	1,017
Panel B: Follow-Up Effects in High School							
Tch. Of Color				0.178*	0.223**	-0.083*	-0.024
				(0.083)	(0.082)	(0.035)	(0.034)
Students				663	670	743	743

Note: All models control for school and grade fixed effects and background teacher characteristics (i.e., gender, certification pathway, teaching experience). Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. Chronically absent and ever suspended variables are binary; other outcomes are standardized.

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Table 7

Effect of Race/Ethnicity Matching on Short- and Longer-Term Student Outcomes

	Self-Efficacy	Engagement/ Happiness in Class	Self-Regulation	Math Achievement	Reading Achievement	Chronically Absent	Ever Suspended
Panel A: End-of-Year Effects in Upper-Elementary School							
Tch. Of Color*Stu. of Color	0.469***	0.307*	0.135	0.242**	0.232*	-0.041	0.017
	(0.120)	(0.143)	(0.117)	(0.089)	(0.100)	(0.027)	(0.028)
Tch. of Color*White Stu.	0.343*	-0.115	0.092	0.431*	0.345~	-0.032	0.104~
	(0.144)	(0.265)	(0.210)	(0.170)	(0.188)	(0.046)	(0.062)
White Tch*White Stu.	0.059	0.029	0.099	0.155*	0.227**	0.008	-0.035*
	(0.081)	(0.101)	(0.079)	(0.076)	(0.068)	(0.027)	(0.016)
P-Value on Tests of Coefficient Equivalence							
Tch. of Color*Stu. of Color = Tch. of Color*White Stu.	0.367	0.124	0.828	0.279	0.587	0.868	0.173
Tch. of Color*White Stu. = White Tch.*White Stu.	0.081	0.605	0.973	0.123	0.551	0.413	0.031
Students	903	903	901	1,003	1,004	1,017	1,017
Panel B: Follow-Up Effects in High School							
Tch. Of Color*Stu. of Color				0.185*	0.195*	-0.091*	-0.027
				(0.080)	(0.090)	(0.038)	(0.034)
Tch. of Color*White Stu.				0.322~	0.493**	-0.011	-0.024
				(0.183)	(0.174)	(0.074)	(0.089)
White Tch*White Stu.				0.356***	0.243*	0.050	-0.030
				(0.098)	(0.093)	(0.042)	(0.037)
P-Value on Tests of Coefficient Equivalence							
Tch. of Color*Stu. of Color = Tch. of Color*White Stu.				0.425	0.121	0.331	0.976
Tch. of Color*White Stu. = White Tch.*White Stu.				0.867	0.184	0.461	0.95
Students				663	670	743	743

Note: All models control for school and grade fixed effects and background teacher characteristics (i.e., gender, certification pathway, teaching experience). Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. Chronically absent and ever suspended variables are binary; other outcomes are standardized. Left-out category is White teacher with student of color.

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Table 8
Differences in Practices and Mindsets Between Teachers of Color and White Teachers in Full Project Sample

	Teachers	(1)	(2)	(4)	(5)
Growth Mindset Beliefs	297	0.235~	0.231~	0.202	0.201
		(0.125)	(0.125)	(0.160)	(0.160)
Relationships with Students and Families	313	0.098	0.096	0.262~	0.299*
		(0.122)	(0.123)	(0.137)	(0.137)
Preparation for Instruction	300	0.616***	0.589***	0.638***	0.647***
		(0.118)	(0.128)	(0.158)	(0.154)
Classroom Support	314	-0.022	0.005	0.132	0.098
		(0.130)	(0.136)	(0.179)	(0.177)
Classroom Organization	314	0.091	0.169	0.245	0.218
		(0.134)	(0.139)	(0.183)	(0.175)
Background Teacher Characteristics			X	X	
School Fixed Effects				X	X

Note: Estimates come from separate regression models. Background teacher characteristics include gender, certification pathway, and teaching experience. All teacher practice and mindset variables are standardized. Heteroskedasticity-robust standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Table 9
Mediating Effects of Teacher Mindsets and Practices on Short- and Longer-Term Student Outcomes

	Self-Efficacy	Engagement/ Happiness in Class	Self-Regulation	Math Achievement	Reading Achievement	Chronically Absent	Ever Suspended
Panel A: End-of-Year Effects in Upper-Elementary School							
Tch. of Color	0.454***	0.244	0.125	0.259**	0.225*	-0.041~	0.036
	(0.117)	(0.149)	(0.116)	(0.086)	(0.089)	(0.024)	(0.028)
Growth Mindset Beliefs	-0.023	0.031	-0.011	-0.008	0.029	0.007	-0.021*
	(0.038)	(0.056)	(0.036)	(0.027)	(0.024)	(0.007)	(0.009)
Tch. of Color	0.414***	0.226	0.123	0.250**	0.241*	-0.026	0.023
	(0.109)	(0.140)	(0.124)	(0.091)	(0.102)	(0.020)	(0.024)
Relationships Stu./Fam.	0.061*	0.045	-0.001	0.011	-0.015	-0.020***	0.013
	(0.026)	(0.033)	(0.035)	(0.028)	(0.026)	(0.006)	(0.008)
Tch. of Color	0.450***	0.258*	0.119	0.260**	0.233*	-0.040	0.030
	(0.115)	(0.127)	(0.115)	(0.084)	(0.094)	(0.025)	(0.026)
Prep. for Instruction	0.023	0.135	-0.065	0.041	0.031	0.002	-0.020
	(0.066)	(0.092)	(0.048)	(0.054)	(0.047)	(0.014)	(0.016)
Tch. of Color	0.449***	0.247~	0.118	0.255**	0.232*	-0.040	0.031
	(0.116)	(0.138)	(0.113)	(0.083)	(0.094)	(0.025)	(0.027)
Classroom Organization	-0.007	0.054	0.066	0.077	-0.045	-0.007	0.016
	(0.051)	(0.060)	(0.069)	(0.052)	(0.055)	(0.012)	(0.011)
Students	903	903	901	1,003	1,004	1,017	1,017
Panel B: Follow-Up Effects in High School							
Tch. of Color				0.190*	0.227**	-0.077*	-0.016
				(0.084)	(0.083)	(0.036)	(0.034)
Growth Mindset Beliefs				-0.035	-0.010	-0.024*	-0.033*
				(0.036)	(0.038)	(0.012)	(0.014)
Tch. of Color				0.148~	0.204*	-0.070*	-0.023
				(0.078)	(0.081)	(0.031)	(0.035)
Relationships Stu./Fam.				0.075**	0.045	-0.023*	-0.002
				(0.027)	(0.032)	(0.009)	(0.012)
Tch. of Color				0.181*	0.221**	-0.084*	-0.024
				(0.084)	(0.079)	(0.035)	(0.033)
Prep. for Instruction				0.033	-0.045	-0.038~	0.001
				(0.064)	(0.064)	(0.020)	(0.028)
Tch. of Color				0.184*	0.222**	-0.083*	-0.024
				(0.078)	(0.082)	(0.034)	(0.033)
Classroom Organization				0.098	-0.035	-0.012	0.015
				(0.066)	(0.058)	(0.024)	(0.021)
Students				663	670	743	743

Note: All models control for school and grade fixed effects and background teacher characteristics (i.e., gender, certification pathway, teaching experience). Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. Chronically absent and ever suspended variables are binary; other outcomes are standardized.
*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Appendices

Appendix Table 1
Item Text from Student Survey

Dimensions and Item Text
Self-Efficacy
I have pushed myself hard to completely understand math in this class
If I need help with math, I make sure that someone gives me the help I need.
If a math problem is hard to solve, I often give up before I solve it.
Doing homework problems helps me get better at doing math.
In this class, math is too hard.
Even when math is hard, I know I can learn it.
I can do almost all the math in this class if I don't give up.
I'm certain I can master the math skills taught in this class.
When doing work for this math class, focus on learning not time work takes.
I have been able to figure out the most difficult work in this math class.
Engagement and Happiness in Class
This math class is a happy place for me to be.
Being in this math class makes me feel sad or angry.
The things we have done in math this year are interesting.
Because of this teacher, I am learning to love math.
I enjoy math class this year.
Self-Regulation
My behavior in this class is good.
My behavior in this class sometimes annoys the teacher. (Reverse coded.)
My behavior is a problem for the teacher in this class. (Reverse coded.)

Appendix Table 2

Pairwise Correlations between Upper-Elementary and High School Outcomes

Upper-Elementary School Outcomes	High School Outcomes					
	Math Achievement	ELA Achievement	Absences	Chronically Absent	Suspensions	Ever Suspended
Intrapersonal Competencies						
Self-Efficacy	0.262***	0.196***	-0.099*	-0.067~	-0.120**	-0.061
Engagement/Happiness in Class	0.180***	0.153**	-0.066~	0.079*	-0.087*	-0.048
Self-Regulation	0.221***	0.272***	-0.116**	-0.107**	-0.163***	-0.163***
Test Scores						
Math Achievement	0.687***	0.560***	-0.178***	-0.141**	-0.148**	-0.117**
ELA Achievement	0.550***	0.636***	-0.132**	-0.126***	-0.142**	-0.105**
School Behaviors						
Absences (days)	-0.168***	-0.069~	0.570***	0.434***	0.071~	0.063~
Chronically Absent	-0.103**	-0.088*	0.475***	0.321***	0.066~	0.030
Suspensions	-0.122**	-0.108**	0.103**	0.133**	0.162***	0.164***
Ever Suspended	-0.132**	-0.129**	0.142**	0.156***	0.210***	0.185***

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Appendix Table 3
Item Text from Teacher Survey

Dimensions and Item Text	Scale
<u>Growth Mindset Beliefs</u>	
The amount my students can learn is primarily related to family background and/or student effort.	We are interested in your ideas about intelligence. There are no right or wrong answers. Please indicate the extent to which you agree or disagree with each of the following statements. 1 = "Strongly Disagree" to 6 = "Strongly Agree" (Reverse Coded)
I am limited in what I can achieve because of student home environment and/or effort.	
Students have a certain amount of intelligence, and I can't do much to change it.	
Students have certain amount of intelligence, and they can't really do much to change it.	
Intelligence is something about students that they can't change very much.	
Students can learn new things, but they can't really change their basic intelligence.	
To be honest, students can't really change how intelligent they are.	
<u>Relationships with Students and Families</u>	
Talking with parents about students' learning or behavior.	In a typical week, how much time do you devote to the following activities? 1 = "No Time" to 5 = "More than Six Hours"
Students and I show an interest in each other's lives.	How often do you observe the following situations while teaching? 1 = "Rarely or Never" to 5 = "Always"
Students and I have a friendly rapport.	
Students and I use respectful language and listen to each other.	
<u>Preparation for Instruction</u>	
Seeking outside support for struggling students in any subject (e.g., IEPs, tutoring).	In a typical week, how much time do you devote to the following activities? 1 = "No Time" to 5 = "More than Six Hours"
Collaboratively planning lessons in any subject with other teachers or coaches.	
Grading mathematics assignments.	
Gathering and organizing mathematics lesson material (e.g., locating and copying supplemental material, preparing manipulatives).	
Reviewing the content of specific mathematics lessons (e.g., reading the teacher manual, seeking additional information about the content).	
Preparing for a mathematics lesson by trying out explanations, or working through examples of problems.	
Helping students learn any subject after school hours (e.g., homework club, tutoring).	
I differentiate mathematics assignments based on students' individualized learning needs.	About how often do you or your students take part in the following activities? 1 = "Never" to 5 = "Daily or Almost Daily"
I evaluate student work on mathematics assessments or assignments using a written rubric.	
I provide detailed written feedback on student mathematical work in addition to a numeric score.	
I examine student work to understand the process students use to solve mathematics problems.	
Students evaluate their own mathematical work on assessments or assignments using a written rubric.	
I change my lesson plans based on what I learn from analyzing student work.	
I design assignments that reveal student thinking rather than just mastery of learning goals.	
I assess students' understanding of a topic before I teach it.	

Appendix Table 4
Item Text from Classroom Assessment Scoring System (CLASS) Observation Instrument

Dimensions	Description
	Classroom Support
Positive Climate	Positive climate reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and non-verbal interactions.
Teacher Sensitivity	Teacher sensitivity reflects the teacher's timely responsiveness to the academic, social/emotional, behavioral, and developmental needs of individual students and the entire class.
Respect for Student Perspectives	Regard for student perspectives captures the degree to which the teacher's interactions with students and classroom activities place an emphasis on students' interests and ideas and encourage student responsibility and autonomy. Also considered is the extent to which content is made useful and relevant to the students.
Instructional Learning Format	Instructional learning format focuses on the ways in which the teacher maximizes student engagement in learning through clear presentation of material, active facilitation, and the provision of interesting and engaging lessons and materials.
Content Understanding	Content understanding refers to both the depth of lesson content and the approaches used to help students comprehend the framework, key ideas, and procedures in an academic discipline. At a high level, this refers to interactions among the teacher and students that lead to an integrated understanding of facts, skills, concepts, and principles.
Analysis and Problem Solving	Analysis and problem solving assesses the degree to which the teacher facilitates students' use of higher-level thinking skills, such as analysis, problem solving, reasoning, and creation through the application of knowledge and skills. Opportunities for demonstrating metacognition, i.e., thinking about thinking, are also included.
Quality of Feedback	Quality of feedback assess the degree to which feedback expands and extends learning and understanding and encourages student participation. Significant feedback may also be provided by peers.
Instructional Dialogue	Instructional dialogue captures the purposeful use of dialogue (structured, cumulative questioning and discussion which guide and prompt students) to facilitate students' understanding of content and language development.
Student Engagement	This scale captures the degree to which all students in the class are focused and participating in the learning activity presented or facilitated by the teacher.
	Classroom Organization
Negative Climate	Negative climate reflects the overall level of negativity among teachers and students in the class.
Behavior Management	Behavior management encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior.
Productivity	Productivity considers how well the teacher manages time and routines so that instructional time is maximized. This dimension captures to degree to which instructional time is effectively managed and down time is minimized for students.

Note: Item text comes directly from the Classroom Assessment Scoring System (CLASS) upper-elementary manual (Pianta, Hamre, & Mintz, 2012).

Appendix Table 5

Availability of Teacher Practices and Mindsets Data and Relationship to Background Teacher Characteristics

	Proportion with Data	P-Value	
		Teacher of Color	Other Teacher Characteristics
Growth Mindset Beliefs	0.86	0.121	0.561
Relationships with Students and Families	0.99	0.666	0.982
Preparation for Instruction	0.86	0.121	0.561
Classroom Observations	0.94	0.638	0.401
Teachers		71	

Notes: P-values test the hypothesis that attrition/missingness is related whether or not a teacher if White or a person of color, or to other background teacher characteristics (i.e., gender, teaching experience, certification pathway).

Appendix Table 6

Effect of Teachers of Color on Short- and Longer-Term Student Outcomes, Including Controls for Background Student and Class Characteristics

	Self-Efficacy	Engagement/ Happiness in Class	Self-Regulation	Math Achievement	Reading Achievement	Chronically Absent	Ever Suspended
				Panel A: End-of-Year Effects in Upper-Elementary School			
Tch. Of Color	0.270**	0.098	0.056	0.163*	0.182**	-0.020	0.072*
	(0.095)	(0.128)	(0.089)	(0.065)	(0.064)	(0.029)	(0.030)
Students	903	903	901	1,003	1,004	1,017	1,017
				Panel B: Follow-Up Effects in High School			
Tch. Of Color				0.145*	0.232**	-0.104***	-0.011
				(0.071)	(0.080)	(0.028)	(0.039)
Students				663	670	743	743

Note: All models control for school and grade fixed effects; background teacher characteristics (i.e., gender, certification pathway, teaching experience); background student (i.e., gender, race/ethnicity, eligibility for free or reduced-price meals and for special education services, limited English proficiency status, and prior-year test scores, absences, and suspensions); and class characteristics that are averages of the student-level variables. Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. Chronically absent and ever suspended variables are binary; other outcomes are standardized.

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Appendix Table 7

Effect of Teachers of Color on Short- and Longer-Term Student Outcomes, Excluding Controls for Background Teacher Characteristics

	Self-Efficacy	Engagement/ Happiness in Class	Self-Regulation	Math Achievement	Reading Achievement	Chronically Absent	Ever Suspended
				Panel A: End-of-Year Effects in Upper-Elementary School			
Tch. Of Color	0.401***	0.246~	0.077	0.216*	0.224*	-0.042~	0.041
	(0.111)	(0.124)	(0.104)	(0.085)	(0.088)	(0.023)	(0.026)
Students	903	903	901	1,003	1,004	1,017	1,017
				Panel B: Follow-Up Effects in High School			
Tch. Of Color				0.103	0.210**	-0.053	-0.039
				(0.079)	(0.070)	(0.039)	(0.031)
Students				663	670	743	743

Note: All models control for school and grade fixed effects. Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. Chronically absent and ever suspended variables are binary; other outcomes are standardized.

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Appendix Table 8
Differences in Practices and Mindsets Between Teachers of Color and White Teachers in Experimental Sample

	Teachers	(1)	(2)	(4)	(5)
Growth Mindset Beliefs	61	0.395	0.437	0.488	0.548
		(0.286)	(0.307)	(0.437)	(0.396)
Relationships with Students and Families	70	0.663**	0.732**	0.411	0.431
		(0.233)	(0.268)	(0.320)	(0.319)
Preparation for Instruction	61	0.388~	0.383	0.018	0.054
		(0.221)	(0.284)	(0.317)	(0.254)
Classroom Support	67	0.202	0.210	0.300	0.278
		(0.296)	(0.328)	(0.384)	(0.377)
Classroom Organization	67	0.507*	0.549**	0.242	0.165
		(0.197)	(0.204)	(0.215)	(0.221)
Background Teacher Characteristics			X	X	
School Fixed Effects				X	X

Note: Estimates come from separate regression models. Background teacher characteristics include gender, certification pathway, and teaching experience. All teacher practice and mindset variables are standardized. Heteroskedasticity-robust standard errors in parentheses.
*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Appendix Table 9
Mediating Effects of Teacher Mindsets and Practices on Short- and Longer-Term Student Outcomes, Conditional on other Mindsets and Practices

	Self-Efficacy	Engagement/ Happiness in Class	Self-Regulation	Math Achievement	Reading Achievement	Chronically Absent	Ever Suspended
				Panel A: End-of-Year Effects in Upper-Elementary School			
Tch. of Color	0.419***	0.242~	0.114	0.265**	0.239*	-0.026	0.024
	(0.120)	(0.136)	(0.119)	(0.093)	(0.108)	(0.020)	(0.022)
Growth Mindset Beliefs	-0.025	0.008	-0.013	-0.024	0.033	0.006	-0.022*
	(0.047)	(0.061)	(0.040)	(0.029)	(0.025)	(0.007)	(0.009)
Relationships Stu./Fam.	0.064*	0.019	0.005	-0.003	-0.016	-0.022***	0.015*
	(0.029)	(0.029)	(0.033)	(0.032)	(0.028)	(0.005)	(0.006)
Prep. for Instruction	-0.000	0.124	-0.063	0.051	0.029	0.011	-0.022
	(0.079)	(0.098)	(0.054)	(0.060)	(0.056)	(0.015)	(0.015)
Classroom Organization	-0.013	0.048	0.069	0.087~	-0.052	-0.003	0.019~
	(0.053)	(0.069)	(0.071)	(0.050)	(0.052)	(0.012)	(0.011)
Students	903	903	901	1,003	1,004	1,017	1,017
				Panel B: Follow-Up Effects in High School			
Tch. of Color				0.179*	0.188*	-0.067*	-0.009
				(0.078)	(0.078)	(0.032)	(0.033)
Growth Mindset Beliefs				-0.050	0.004	-0.022~	-0.039**
				(0.038)	(0.038)	(0.012)	(0.015)
Relationships Stu./Fam.				0.062*	0.063*	-0.020*	-0.008
				(0.028)	(0.031)	(0.009)	(0.011)
Prep. for Instruction				0.018	-0.073	-0.024	0.013
				(0.070)	(0.067)	(0.023)	(0.028)
Classroom Organization				0.095	-0.058	0.003	0.032
				(0.064)	(0.059)	(0.024)	(0.020)
Students				663	670	743	743

Note: All models control for school and grade fixed effects and background teacher characteristics (i.e., gender, certification pathway, teaching experience). Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. Chronically absent and ever suspended variables are binary; other outcomes are standardized.

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1

Appendix Table 10

Mediating Effects of Teacher Mindsets and Practices on Short- and Longer-Term Student Outcomes, Replacing Randomization Block Fixed Effects with Observable School Characteristics

	Self-Efficacy	Engagement/ Happiness in Class	Self-Regulation	Math Achievement	Reading Achievement	Chronically Absent	Ever Suspended
				Panel A: End-of-Year Effects in Upper-Elementary School			
Tch. of Color	0.334***	0.166	0.101	0.249*	0.221*	-0.033	0.041~
	(0.090)	(0.136)	(0.105)	(0.109)	(0.098)	(0.021)	(0.022)
Growth Mindset Beliefs	-0.023	-0.001	-0.000	0.004	0.042	0.006	-0.022*
	(0.037)	(0.053)	(0.038)	(0.038)	(0.032)	(0.008)	(0.009)
Tch. of Color	0.290***	0.136	0.099	0.245*	0.240*	-0.020	0.029
	(0.083)	(0.133)	(0.113)	(0.112)	(0.107)	(0.017)	(0.020)
Relationships Stu./Fam.	0.076**	0.059~	0.003	0.008	-0.017	-0.020**	0.012
	(0.027)	(0.035)	(0.037)	(0.032)	(0.028)	(0.006)	(0.008)
Tch. of Color	0.327***	0.147	0.110	0.240*	0.224*	-0.032	0.038
	(0.086)	(0.125)	(0.106)	(0.102)	(0.100)	(0.021)	(0.023)
Prep. for Instruction	0.010	0.131~	-0.065	0.101~	0.050	-0.001	-0.013
	(0.040)	(0.066)	(0.045)	(0.058)	(0.047)	(0.013)	(0.010)
Tch. of Color	0.324***	0.145	0.086	0.233*	0.226*	-0.030	0.035
	(0.088)	(0.127)	(0.104)	(0.105)	(0.102)	(0.021)	(0.023)
Classroom Organization	0.023	0.091	0.067	0.100~	0.019	-0.007	0.008
	(0.052)	(0.068)	(0.054)	(0.052)	(0.053)	(0.013)	(0.014)
Students	903	903	901	1,003	1,004	1,017	1,017
				Panel B: Follow-Up Effects in High School			
Tch. of Color				0.143	0.226**	-0.068*	-0.029
				(0.093)	(0.077)	(0.034)	(0.027)
Growth Mindset Beliefs				-0.033	-0.007	-0.028*	-0.033*
				(0.041)	(0.040)	(0.013)	(0.015)
Tch. of Color				0.108	0.209**	-0.062~	-0.036
				(0.084)	(0.077)	(0.034)	(0.031)
Relationships Stu./Fam.				0.064*	0.037	-0.024*	-0.001
				(0.026)	(0.032)	(0.009)	(0.011)
Tch. of Color				0.134	0.226**	-0.073*	-0.037
				(0.090)	(0.074)	(0.033)	(0.030)
Prep. for Instruction				-0.004	-0.050	-0.039*	0.005
				(0.052)	(0.058)	(0.019)	(0.023)
Tch. of Color				0.127	0.226**	-0.071*	-0.038
				(0.084)	(0.077)	(0.033)	(0.030)
Classroom Organization				0.086	-0.015	-0.036	0.010
				(0.060)	(0.049)	(0.024)	(0.023)
Students						743	743

Note: All models control for fixed effects for grade level, background school characteristics, and background teacher characteristics (i.e., gender, certification pathway, teaching experience). Heteroskedasticity-robust standard errors clustered at the teacher level in parentheses. Chronically absent and ever suspended variables are binary; other outcomes are standardized.

*** p<0.001, ** p<0.01, * p<0.05, ~ p<0.1