"Model Minorities" in the Classroom? Positive Evaluation Bias towards Asian Students and its Consequences*

Ying Shi
Syracuse University†

Maria Zhu
Syracuse University‡

Abstract

The fast-growing demographic group of Asian Americans is often perceived as a "model minority." This paper establishes empirical evidence of this stereotype in the context of education and then analyzes its consequences. We show that teachers rate Asian students' academic skills more favorably than observationally similar White students in the same class, even after accounting for test performance and behavior. This contrasts with teachers' lower likelihood of favoring Black and Hispanic students. Notably, teachers respond to the presence of any Asian student in the classroom by exacerbating Black-White and Hispanic-White assessment gaps. This suggests that the "model minority" stereotype can negatively impact other minority groups despite its ostensibly positive connotation.

JEL: I24, J15

Keywords: Teacher evaluation, racial bias, Asian Americans

---

*We are grateful to Peter Arcidiacono, Elizabeth Cascio, Katherine Harris-Lagoudakis, Jennifer Lee, Soumaya Keynes, Jessica Merkle, Emily Owens, Natasha Warikoo, Emily Weisburst, Arezou Zaresani, and conference participants at AEFP, APPAM, NBER Race and Stratification Working Group, and SEA for helpful comments. Declaration of interest: none.

†Corresponding author. Shi: Department of Public Administration and International Affairs, 426 Eggers Hall, Syracuse, NY 13244-1020 (Email: yshi78@syr.edu; Phone: 315-443-9442)

‡Zhu: Department of Economics, 110 Eggers Hall, Syracuse, NY 13244-1020 (Email: mzhu33@syr.edu; Phone: 315-443-9043)

1 Introduction

Asian Americans currently represent the single fastest growing racial and ethnic group in the United States (Budiman 2020). They experience a unique profile of racial stereotypes compared to other minority groups in the country. Since the mid-1900s, Asian Americans have been lauded as the nation's "model minority," due to perceived success in assimilation, upward mobility, and educational achievement (Wu 2014). The view of Asians as "model minorities" is pervasive in education given their ability to outperform other racial and ethnic groups on standardized tests and grades on average (Fejgin, 1995; Hsin & Xie, 2014; Kao, 1995) and record of postsecondary enrollment and attainment at selective institutions (Sakamoto, Goyette, & Kim, 2009).

On one hand, this view of Asian Americans confers benefits through "stereotype promise," in which being viewed through positive lens may enhance performance for the positively stereotyped group (Lee & Zhou, 2015). On the other hand, while this "positive" stereotype is ostensibly beneficial, there is concern that it could carry negative consequences. For example, it may hold individuals in the stereotyped groups to unrealistically high expectations (C. Ho, Driscoll, and Loosbrock 1998), hinder their performance (Cheryan and Bodenhausen 2016), or constrain stereotyped group members in their pursuit of certain academic and career tracks (Czopp 2010). There also may be negative effects if positive stereotypes for Asians reinforce the notion of fundamental differences across groups or bolster negative stereotypes for other, under-represented minority groups (Kay, Day, Zanna, and Nussbaum 2013).

This study provides evidence on the presence and consequences of positive teacher assessment bias towards Asian students in schools. Racial biases in teachers' assessments are important to understand, given the role that teacher evaluations play in several educational domains, including recommending students for various academic tracks, writing letters of recommendation, and conveying to students what is expected from them. We first examine whether teachers systematically evaluate Asian students differently than same-class White peers with the same performance, before exploring how the magnitude of any assessment differentials vary across Asian ethnic subgroups. Finally, this paper analyzes whether the propensity of teachers to favor Asian students has spillover effects, by examining how teachers change their assessments of other, under-represented minority groups in the presence of an Asian student in the classroom.

To address our research questions, we use administrative data from the North Carolina Education Research Data Center (NCERDC) covering students in grades 3-8. The NCERDC dataset has several key advantages that make it uniquely well-suited for this study. First, the data contain two different measures of a student's academic ability: standardized test scores and teacher assessments. The juxtaposition of these measures, which are designed to capture the same underlying skillset, enables the identification of teacher bias. Second, the data provide a significant number of Asian students for analysis because it spans all public elementary and middle school students in North Carolina from 2007 to 2013. While there is a sizable number of Asian students in our sample, they still comprise a small enough share of overall enrollment that having an Asian student in class is not a regular occurrence. As a result, we are able to assess the nature of spillover effects arising from the modal experience of exposure to a single Asian student, which would be more difficult to do in settings with high shares of Asian students.

Both teacher assessments and standardized test scores in math and reading are mapped onto a discrete 1 to 4 scale, which allows us to directly compare these two measures of achievement. We use standardized test score-based achievement levels to provide a benchmark for assessing whether teachers are systematically over-rating or under-rating Asian students relative to other groups, conditional on student achievement and a rich vector of individual sociodemographic and behavioral attributes. We observe a student's raw, uncoarsened standardized test scores, which allows us to flexibly control for academic performance. In addition to these controls, our analyses also include classroom-level fixed effects to address any endogeneity in teacher evaluations that could arise at the teacher, year, school, subject, and grade level.

Results indicate that teachers display significant positive bias towards Asian students, relative to White students in the same class with the same standardized test scores and sociodemographic and behavioral characteristics. Compared to White students, teachers are 4.3 percentage points more likely to give Asian students a higher evaluation (over-rate) than the blind-scored achievement level indicated by their standardized test scores and 2.7 percentage points less likely to give Asian students a lower evaluation (under-rate). These magnitudes correspond to 12 percent and 14 percent of baseline propensities to over-rate and under-rate students, respectively, indicating that teachers' propensities for favoring Asian students are sizable. We perform several robustness checks to rule out alternative explanations for these racial differences, including accounting for the roles of measurement error, hard-to-observe behavioral attributes, differences in assessment standards across classes, and racial biases in standardized testing. These effects are sizable and present in math and reading and in both elementary and middle schools, suggesting positive bias towards Asian students is pervasive across subjects and grade levels. Additionally, we find heterogeneous effects by more fine-grained ethnic subgroups. Teachers display greater positive bias towards Asian students from East and South Asian backgrounds, relative to students from Southeast Asian backgrounds.

Next, our findings suggest that the presence of a single Asian student carries potential negative spillovers to teachers' assessments of other students of color. Specifically, teachers decrease their propensity to over-rate a Black or Hispanic student relative to a White student with the same test scores when there is an Asian student in the classroom, compared to classrooms without any Asian students. We similarly find a significant increase in the propensity for teachers to under-rate Black students when an Asian student is present in the classroom. These findings support the notion that teachers may amplify existing negative biases towards under-represented minority groups in the presence of Asian students and associated positive stereotypes, resulting in cumulative disadvantage for Black and Hispanic students. Notably, these effects are driven by classrooms featuring Asian students whose high achievement adheres to the "model minority" characterization. Teachers' exposure to high-performing Black and Hispanic students does not lead to analogous consequences, suggesting that negative spillovers are a distinct effect of teachers' exposure to stereotype-conforming Asian students.

This paper makes several contributions to existing research. First, it provides empirical evidence on a fast-growing and understudied demographic group, Asian Americans. Despite the rapid growth of Asian Americans as a share of the population, scholarship on their educational and labor market trajectories is still limited in disciplines such as economics and sociology (Altonji & Blank, 1999; Sakamoto et al., 2009). In economics, studies utilizing different datasets, methods, and timelines show Asian Americans attaining varying degrees of earnings parity with their non-Hispanic White counterparts (D. A. Black, Haviland, Sanders, & Taylor, 2008; Chiswick, 1983; Duleep & Sanders, 1992; Mar, 2005; Weinberger, 1998). Increasingly, the evidence points to discrimination as a source of downward pressure on Asian American wages and salaries (Duleep & Sanders, 1992; Hilger, 2017; Mar, 2005).[1] Despite the evidence on labor market discrimination, there is less documentation of potential differential treatment of Asian Americans during the schooling.[2] One exception is Chang and Sue (2003), which

[1]Duleep and Sanders (1992) find that on average, American-born Asian men in the 1980 Census earn the same as their White counterparts, but the relative wages of these Asian men fall after conditioning on occupation and industry in a manner that is consistent with some discrimination against these highly-educated employees. Asian American men are also less likely to be in managerial positions, a finding on the so-called "glass ceiling" that is echoed by Mar (2005). Hilger (2017) shows that the upward mobility of Asian Americans is driven primarily by earning gains conditional on education that reflects declining discrimination in the latter half of the twentieth century. Duleep and Sanders (2012) provides evidence that the Civil Rights Act led to a decline in anti-Asian discrimination that contributed to these labor market shifts. Note that given the wide range of data, methods, and models, some studies do not find evidence of discrimination or glass ceilings (see, for example, Sakamoto, Woo, and Yap (2006)).

[2]More recently, Arcidiacono, Kinsler, and Ransom (2020) have focused on discrimina-

uses a vignette study to assess racial differences in teachers' assessments of behavioral characteristics. The authors find that stereotypes influence teachers' perceptions for Asian American students, especially regarding expectations of overcontrolled behavioral traits, which are viewed as more typical for Asian students.[3] We add to this literature in providing evidence for racial differences in teacher assessments of academic achievement that *favor* Asians relative to White students, in a manner that sets Asian students apart from other, under-represented minority groups. This lends empirical credence to the existence of positive stereotypes.

Notably, the patterns for Asians belie substantial heterogeneity, with diminished positive bias towards Asians from particular ethnic groups (e.g., individuals from Southeast Asian backgrounds) and Asians in urban settings. These findings underscore the need to shift away from a view of Asian Americans as a monolithic group towards one that accommodates a diversity of Asian demographic characteristics and experiences (Chiswick, 1983; Duleep & Sanders, 1992; Lee & Zhou, 2015; Sakamoto et al., 2009; Xie & Goyette, 2004).[4]

In addition to documenting the magnitude of Asian-White teacher rating gaps, we examine how perceptions of Asians interact with teacher ratings of other racial and ethnic groups. Potentially detrimental consequences of teachers' positive bias include the reinforcement of beliefs that there exist fundamental differences between groups and a subsequent increase in the usage of *negative* stereotypes (Kay et al., 2013). Our findings that gaps in teachers' assessments between Black and White students and Hispanic and White students widen after exposure to an Asian student in the same classroom are consistent with a theoretical conception of stereotypes rooted in representativeness (Kahneman and Tversky 1972, Bordalo, Coffman, Gen-

tory behaviors that Asian students face relative to White counterparts in the college admissions process.

[3]Overcontrolled behavioral traits refer to behavior patterns of excessive self-control, such as perfectionistic behavior and rigidity.

[4]Proponents of the demographic heterogeneity approach argue for a disaggregation of Asian Americans into more nuanced categories due to differences in access to resources that may shape labor market trajectories (Sakamoto et al., 2009; Xie & Goyette, 2004).

naioli, and Shleifer 2016), or the frequency in which a type occurs in a group relative to baseline. If Asian students are perceived as high-achievers under the "model minority" stereotype, their presence may increase the application of negative stereotypes toward other, under-represented minority groups.

Finally, this paper contributes to a growing body of research on the role of teacher expectations as an input into education production. A burgeoning literature shows that teacher expectations can vary by student attributes such as race (Burgess & Greaves, 2013; Lavy, 2008; Ouazad, 2014; Rangel & Shi, 2020) and gender (Lavy, 2008; Lindahl, 2016).[5] While papers increasingly document discrepancies in teacher expectations across racial and ethnic groups, there is still scarce research investigating bias towards Asians.[6] Instructor expectations matter because they affect student grades and the steering of students towards academic tracks such as gifted and talented programs and advanced coursework (Donovan and Cross 2002 , Francis 2012, Lindahl 2016, Card and Giuliano 2016, Francis, de Oliveira, and Dimmitt 2019). Students may also adjust their behaviors and academic trajectories in ways that render teacher expectations as self-fulfilling prophecies (Rosenthal and Jacobson 1968, Ouazad and Page 2013, Jussim and Harber 2016, Lavy and Sand 2018, Lavy and Megalokonomou 2019, Papageorge, Gershenson, and Kang 2020, Hill and Jones 2021). The consequences of teacher expectations endure through postsecondary education in some instances (Papageorge et al. 2020) but are less persistently documented in others (Hill and Jones 2021).[7]

[5]The interaction between teacher and student attributes matters, as congruence in race, gender, or immigration status can manifest in more favorable teacher assessments (Lindahl, 2016; Ouazad, 2014).

[6]An exception is Burgess and Greaves (2013), which juxtaposes teacher assessments in the English testing system across Asian subgroups such as Indian, Chinese, Bangladeshi, and Pakistani.

[7]Hill and Jones (2021) and Papageorge et al. (2020) use different contexts and identification strategies to examine the impact of differential teacher assessments. The former uses an instrumental variables strategy with a rich set of fixed effects for elementary and middle school students, while the latter relies on within-student variation in tenth-grade teacher expectations. Hill and Jones (2021) find that teacher evaluations matter for student perfor-

In the remainder of the paper, Section 2 presents our data and provides an in-depth overview of the blind and non-blind evaluation measures used in the paper. Section 3 discusses the empirical strategy used to identify differences in teacher evaluations across student race. Section 4 presents our results and Section 5 concludes.

2 Data and Descriptive Statistics

2.1 North Carolina Education Data

This study uses statewide administrative records from the North Carolina Education Research Data Center (NCERDC). Student-level data contain socio-demographic information on gender, race and ethnicity, and economic disadvantage status. The NCERDC also reports individuals' primary home language, which we use as a proxy to inform more detailed information on students' ethnicities and countries of origin.

Similarly, we observe teacher-level attributes including race, ethnicity, and age. Longitudinal data on when a teacher was first observed in a North Carolina traditional public or charter school allow us to determine teachers' years of experience. Detailed course membership rosters with unique student and teacher IDs enable the linking of student sociodemographic data with teacher records and course attendance. We focus on students in grades 3-8 from 2007-2013, which is the sample for which we observe both course membership and teacher assessment information.

An important feature of the data is the presence of both blind-scored assessments and non-blind teacher evaluations of student performance along the same scale. Students take End-of-Grade (EOG) standardized tests in math and reading from third through eighth grade. These tests are given during the last three weeks of the school year, with questions formulated in

mance, particularly for earlier grades, although these effects do not persist. Papageorge et al. (2020) document more persistent causal effects through college completion. The mixed evidence on the enduring effects of teacher expectations on student outcomes is consistent with reviews of the literature in social psychology and beyond (Jussim & Harber, 2005).

a multiple-choice format. Raw student scores on EOG tests are mapped to achievement levels on a discrete scale from 1 to 4 denoting score cutoffs relative to grade-level comparisons. We observe in the data both raw EOG test scores, as well as the 1 to 4 achievement level the raw test score maps to. Levels 1 to 4 refer to insufficient mastery, inconsistent mastery, consistent mastery, and superior performance, respectively.[8] We refer to standardized test assessments of math and reading ability as "blind" assessments, since EOG tests are machine-scored, without regard to a student's identity.

Teacher evaluations map to the same four-point scale of achievement levels for each subject. We refer to teacher assessments of students as "non-blind" assessments since teachers inevitably need to know the identity of the student in question in order to evaluate the student. With knowledge of a student's identity comes information about and the race and ethnicity of each student. We examine whether this information influences how teachers perceive a student's skill-based achievement level.

2.2 Teacher Evaluations

Teacher evaluations of student skills in math and reading come from End-of-Grade data files. Concurrently with the state administration of EOG exams, teachers are asked to provide their assessment of each student's skill mastery on the four-point achievement level scale corresponding to insufficient, inconsistent, consistent, or superior mastery. Given the timing, teach-

[8] A detailed description of each achievement level is as follows:

1. Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.
2. Students performing at this level demonstrate inconsistent mastery of knowledge and skills in this subject area and are minimally prepared to be successful at the next grade level.
3. Students performing at this level consistently demonstrate mastery of grade level subject matter and skills and are well prepared for the next grade level.
4. Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level work

ers submit evaluations *before* they observe students' end-of-year standardized test results.

There is one stated reason for asking teachers for these evaluations. The state uses the average of these teacher judgments to calibrate cut points in the aforementioned four-point scale. Teacher assessments are only one input, as the state also takes into consideration expert input and standard-setting processes. These assessments are not used for any other purpose, such as teachers' performance evaluations. This implies that teachers lack incentives to misrepresent their assessments of student performance.

In order to interpret racial differences in teacher assessments for students in a given classroom with comparable EOG performance as evidence of bias, we need to establish that teacher ratings aim to measure the same underlying skills as EOG tests. We advance several reasons for a close correspondence in content between these two types of assessments. First, the questionnaire instructions for student evaluations explicitly ask teachers to focus their evaluation on the tested subject. As such, the sequence of signals the teacher receives about a student's science competence should not be an input into their assessment of math mastery, or vice versa. Second, teachers were asked to evaluate students' "absolute" ability. This means that teachers are not judging student performance relative to peers in the same classroom or school, but rather to a common statewide standard that is external to the test. The four-point achievement scales used in teacher and EOG assessments align closely with the North Carolina Standard Course of Study, which defines the curriculum standards for each grade and subject to ensure uniformity across classrooms statewide. Teachers undergo training on standards-based grading to minimize subjectivity, thereby enhancing familiarity with state-defined standard objectives. They furthermore have access to the descriptions of skills associated with each achievement level when they evaluate students.

Accountability pressures also induce teachers to spend greater time preparing their students for standardized exams. To the extent that teachers use practice EOG tests or similar materials, students' aptitude on these assessments likely serve as inputs into both teacher evaluations and the actual EOG test, thereby strengthening their relation to each other. Finally, teachers are explicitly instructed to assess students based on achievement, rather than behavior.[9] This further strengthens the relationship between teacher evaluations and achievement-based EOG scores by minimizing the extent to which teachers consider behavioral or socioemotional factors.

2.3 Descriptive Statistics

The top panel of Table 1 describes our student sample. Approximately 3 percent of students are Asian, while the majority of students (54 percent) are White. One advantage of the NCERDC data is that even though Asians constitute a relatively small proportion of the overall student body, there are still over 40,000 Asian students in our sample to allow for sufficient statistical power. Additionally, the small share of Asians in the student body renders teachers more likely to be exposed to a single Asian student in their classroom, and makes this an especially suitable setting for studying spillover effects on other under-represented groups. Black and Hispanic students make up 27 percent and 12 percent of the sample, respectively. In our main analysis, we use an indicator for economic disadvantage and the lagged number of days absent in a year as a proxy for behavioral differences that may emerge in the classroom. On average, half of the students in this sample are economically disadvantaged, and students were absent for about 7 days in a given school year.

The relatively small share of Asians in the North Carolina administrative data prompts questions on their distribution, in particular whether they are concentrated in specific classrooms. Figure 1 shows that apart from the 73 percent of classrooms with no Asian students, the modal case in 17 percent

[9] The prompt given to teachers reads: *"The [subject] teacher should base this response for each student solely on mastery of [subject]. The [subject] teacher may elect to use grades as a starting point in making these assignments. However, grades are often influenced by factors other than pure achievement, such as failure to turn in homework. The [subject] teacher's challenge is to provide information that reflects only the achievement of each student in the subject matter tested."*

Table 1: Student and Teacher Characteristics

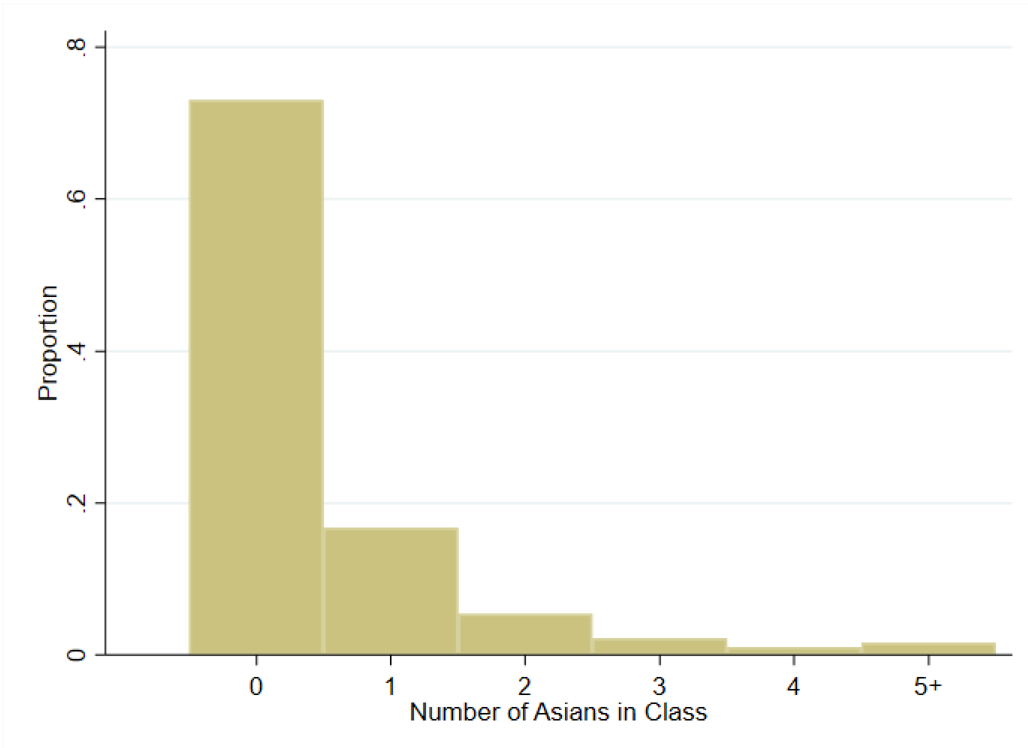
	Mean
Students	
White	0.54
Black	0.27
Hispanic	0.12
Asian	0.03
Native American	0.01
Other race	0.04
Female	0.49
Economically disadvantaged	0.50
Lagged days absent	6.53
	(5.65)
N	1,410,653
Teachers	
White	0.82
Black	0.15
Hispanic	0.01
Asian	0.01
Other race	0.01
Female	0.88
Years of experience	10.39
	(9.67)
N	50,215

Observations in the. top panel are at the student level for students in grades 3-8 in math or reading classes between 2007-2013. A student's lagged number of days absent and status as economically disadvantaged are calculated as the average value of that variable for each year they appear in the data. Observations in the bottom panel are at the teacher level for teachers teaching grades 3-8 in math or reading classes between 2007-2013. Teacher experience is calculated as the average number of years of experience over the period the teacher appears in the data.

of classrooms is one Asian student.

The bottom panel of Table 1 details the characteristics of teachers in the

Figure 1: Asian Representation across Classes



Observations are at the classroom level. The histogram shows the distribution of number of Asians in a classroom.

sample. Relative to students, teachers are disproportionately White (82 percent of the sample). Most of the remaining teachers are Black, and Asians comprise only one percent of the teacher sample. Nearly nine out of every ten teachers are female, a proportion in keeping with national statistics of the elementary and middle school teaching workforce that skews heavily towards women. On average, teachers in our sample period have 10.4 years of experience.

To give a sense for how the academic achievement of Asians compares to other students, the top panel of Table 2 shows the mean raw EOG z-score results by race and corresponding blind-scored achievement levels. The average z-score for the full sample is nearly centered at 0, with a mean achievement level of 2.8. White and Asian students score 0.28 and 0.46 standard deviations above the state average while Black and Hispanic students are approximately one-half and one-third standard deviations below the average, respectively.[10] The bottom panel displays the share of blind-scored achievement levels by racial and ethnic group. Overall, 22 percent of students rank in the top achievement category, level 4. Another 47 percent of students score at level 3, which represents the plurality of students. Compared to both White and under-represented minority students, Asian students have significantly higher average achievement levels and are disproportionately represented in the higher achievement categories. The difference in achievement scores between White and Asian students is concentrated at the top of the distribution. In our sample, 40 percent of Asian students have an achievement level of 4, compared to only 31 percent of White students.

Table 2: Test Scores and Achievement Levels by Race

	All	White	Asian	Black	Hispanic
Mean Raw EOG (z-score)	-0.01	0.28	0.46	-0.48	-0.33
Mean Achievement Level	2.80	3.03	3.12	2.42	2.52
Share of Students at Achievement:					
Level 4	0.22	0.31	0.40	0.09	0.12
Level 3	0.47	0.48	0.39	0.44	0.45
Level 2	0.20	0.14	0.13	0.29	0.26
Level 1	0.11	0.06	0.08	0.19	0.17
N	16,004,741	8,639,535	389,432	4,185,749	1,893,326

Observations represent blind-graded, standardized test scores in math and reading for students from 2007-2013. Two-sample t-test results indicate the mean blind-scored achievement of Asians is significantly larger from that of each of the other racial groups at a 99 percent confidence level.

Table 3 brings in data on teacher ratings on the same four-point scale and juxtaposes these ratings with blind-scored achievement levels based on standardized test scores. Rows denote a student's blind-scored achievement level, and columns represent the teacher rating for the student. Cells

[10]Figure A1 provides additional context by showing the performance distributions of Asian and White students.

denote the proportion of students at each teacher-rated level, conditional on a given blind-scored achievement level. Dark (light) shaded areas denote cells for which a teacher over-rates (under-rates) a student relative to their blind-scored achievement levels.

Table 3: Blind-scored Achievement Levels vs. Teacher Ratings

Blind-scored	Teacher rating							
	White students				**Asian students**			
Achievement	Level 1	Level 2	Level 3	Level 4	Level 1	Level 2	Level 3	Level 4
Level 1	0.22	0.45	0.31	0.02	0.22	0.42	0.32	0.04
Level 2	0.08	0.34	0.50	0.07	0.07	0.29	0.51	0.13
Level 3	0.02	0.15	0.58	0.26	0.02	0.12	0.50	0.36
Level 4	0.00	0.03	0.32	0.65	0.00	0.02	0.24	0.74

Table aggregates math and reading evaluations. Cells represent the share of students who got a blind-score in the row value that were evaluated by their teachers at the column value. Dark (light) shaded areas denote cells for which a teacher over-rate (under-rate) a student relative to their blind-scored achievement levels.

Values in Table 3 indicate teachers may be more likely to over-rate Asians and less likely to under-rate Asians relative to White peers. These patterns are especially stark for high-achieving students, as measured by blind-scored achievement levels. For example, while 26 percent of White students who have a blind achievement level of 3 are rated at an achievement level of 4 by their teachers, this proportion is 36 percent for Asian students. Overall, teachers are eight percentage points more likely to over-rate an Asian relative to a White student, relative to a baseline probability of over-rating among White students of 34 percent. Teachers are four percentage points less likely to under-rate Asian students, relative to a baseline probability of under-rating among White students of 19 percent. Two-sample t-tests reveal that the probability of a teacher to over-rate or under-rate an Asian student differs significantly from their propensity to do so for a White student at a 99 percent confidence level.[11]

While Table 3 provides suggestive evidence that teachers may exhibit

[11]We exclude students with a blind score of 4 in the measurement of over-rating and students who score of 1 in the measurement of under-rating since these students mechanically cannot be over-rated or under-rated.

positive bias towards Asian students relative to White students, these numbers should not be interpreted as causal because they do not control for any underlying differences between White and Asian students themselves or differences in factors affecting their assignment to particular schools, teachers, and classes that may affect assessment scores. The next section discusses in detail potential endogeneity concerns of causal interpretations of these correlations and presents the empirical strategy used to identify the presence of teacher biases in student evaluation.

3 Empirical Strategy

Cross-tabulations of subjective teacher assessments and blind-scored standardized test outcomes are unlikely to reflect teacher bias without adjusting for precise student ability, behavior, and conditions governing the assignment of students into classrooms. Our main specification accounts for these factors by estimating the following linear probability model:

$$O_{ic} = \mathbf{R}'_{ic}\beta + \alpha f(E_{ic}) + \mathbf{X}'_{ic}\Omega + \eta_c + \epsilon_{ic} \quad (1)$$

where O_{ic} represents the outcome of interest for student i in class c. We examine two different outcomes: whether the teacher's non-blind ($NB \in \{1,2,3,4\}$) rating is *higher* or *lower* than the student's blind-scored ($B \in \{1,2,3,4\}$) achievement level on the same four-point scale based on standardized test performance. O_{ic} is then expressed as $\mathbb{1}\{NB > B\}$ or $\mathbb{1}\{NB < B\}$, respectively. Students who score a 4 are not included in the over-rating sample since it is mechanically infeasible to over-rate these students. Analogously, students who score a 1 are excluded from the under-rating sample.

This regression framework addresses multiple potential confounding factors in order to isolate racial differences in assessment attributed to teacher bias (as captured by the coefficient on student race indicators $\mathbf{R}'_{ic}$). First, Equation 1 flexibly controls for a student's End-of-Grade exam score, E_{ic}, using subject-, year-, and grade-specific score fixed effects. Specifically, E_{ic}

denotes a student's raw, uncoarsened achievement score that has not been binned or aggregated. These fixed effects control for the common role that underlying achievement plays across race. They also address the possibility that student score distributions within each of the four achievement levels may vary by race.[12]

The vector $\mathbf{X}'_{ic}$ controls for a set of observable characteristics, including student gender, lagged counts of number of days absent, lagged counts of disciplinary incidents or office referrals, and whether the student is economically disadvantaged. These variables address the possibility that student compositions along these characteristics differ across racial groups, which may subsequently affect teacher assessments. In particular, if there are unobserved behavioral components that affect assessment, this may be captured by the lagged absence and disciplinary measures.

Finally, the addition of a class fixed effect, η_c, means identification comes from *within-classroom* variation in teacher assessments. The fixed effect accounts for the possibility that Asian students are disproportionately concentrated in classrooms with more- or less-lenient teachers relative to White counterparts. It also accounts for any classroom-specific shocks that may affect learning, as well as changes across testing standards over time.

To determine how teachers' propensities to over-rate or under-rate differ across student racial and ethnic groups, we examine the coefficient of interest β on the vector of student race and ethnicity indicators ($\mathbf{R}_{ic}$), using White students as the reference category. β captures racial differences in teachers' subjective evaluations within a given class, after adjusting for students' raw standardized test scores and behavioral proxies. We interpret this differential as indicative of teacher racial bias in assessments. An important caveat is that, since teacher assessments take place at the end of the academic year, this definition of racial bias is not inclusive of the effects of

[12]For example, suppose White students who get categorized under achievement level 4 have raw End-of-Grade test scores just above the achievement level 4 threshold, while Asian students with the same achievement level are clustered well above the cutoff. In this scenario, differences in teacher assessments relative to achievement levels may reflect actual differences in achievement, rather than teacher racial biases.

teacher expectations on the contemporaneous test scores used as an input in our models. Differential expectations by student racial group can influence teachers' pedagogy and effort in ways that directly influence student learning, with literature empirically documenting the existence of self-fulfilling prophecies (Hill & Jones, 2021; Papageorge et al., 2020). To the extent this holds in our context, we may be under-estimating the scope of racial biases.

Next, we augment our empirical specification to test for spillover effects of exposure to any Asian students in the classroom. As before, the outcome variable O_{ic} denotes whether the teacher is over-rating ($\mathbb{1}\{NB > B\}$) or under-rating ($\mathbb{1}\{NB < B\}$) student i in classroom c:

$$O_{ic} = \mathbf{R}'_{ic}\pi + (\mathbf{R}'_{ic}\text{AnyAsian}_c)\Phi + \mathbf{R}'_{ic}\delta_j + \rho f(E_{ic}) + \mathbf{X}'_{ic}\Gamma + \theta_c + \epsilon_{ic} \quad (2)$$

The above model follows Equation 1 in flexibly controlling for the student's blind-scored test performance using subject-, year-, and grade-specific score fixed effects, alongside individual attributes such as lagged days absent and disciplinary infractions, economic disadvantage, and gender. The inclusion of a classroom fixed effect, θ_c absorbs classroom-level shocks such as shared disruptions to learning and teacher preferences for grading that are common to all students.

This specification departs from the base model in the inclusion of an interaction term between student race and whether there is at least one Asian student in the classroom (AnyAsian$_c$). Since it is highly plausible that classroom racial composition relates to school and teacher characteristics due to the sorting of students into classrooms, we also include a full set of student race indicators interacted with teacher-school-grade-course fixed effects (δ_j). These absorb fixed differences in the likelihood of having at least one Asian student across teachers in a given school and course type (e.g. fifth grade math at Sycamore Creek Elementary School). The residual variation in AnyAsian$_c$ is then within teacher and course.[13] We infer a causal

[13]Note that 81 percent of schools in our sample have a teacher-school-grade-course cell

interpretation of the parameters of interest, Φ, as the effect of exposure to any Asian student on teacher assessments of other racial groups (relative to White peers), compared to classrooms without Asian students. While we examine spillover effects across all documented racial groups, we focus in particular on Black-White and Hispanic-White assessment gaps. In heterogeneity analyses, we augment the model with interactions between race, the indicator for at least one Asian student, and attributes of those Asian students such as academic achievement or socioeconomic status.

Our empirical strategy assumes idiosyncratic variation in exposure to at least one Asian student for a teacher in a given school, grade, and course.[14] We advance that this is a plausible assumption given natural population variation in the presence of students of a particular racial or ethnic group. We also restrict the analytic sample to only classrooms with zero or one Asian student so that results are not identified off of classrooms with larger numbers of Asian students.[15] To further assess the validity of our assumption, we examine the relationship between having one Asian student and class-level characteristics including gender, racial and ethnic, and socioeconomic composition, as well as achievement score gaps. While these at-

that exhibits some variation in having an Asian student, while 19 percent of schools have zero cells with any variation. The schools for which we can identify spillover effects are less prevalent in rural areas with somewhat smaller student populations.

[14]Our reliance on cross-cohort variation in the number of students belonging to a particular racial group recalls other papers in the peer effects literature (see, for example, Bifulco, Fletcher, and Ross (2011) and S. Black, Devereux, and Salvanes (2013)). We depart from other studies by using variation at the finer level of teacher, course, school, and grade cells.

[15]There are several reasons for this sample restriction. First, Figure 1 documents that the modal case for any exposure is to a single Asian student, with classrooms having up to one Asian student making up 90 percent of the sample. The second is that restricting to classrooms with no more than one Asian student facilitates the interpretation of coefficients on race variables interacted with AnyAsian_c and attributes such as the student achievement. When there are multiple Asian students, we need to make assumptions about the signal that teachers extract from these students who are performing at different levels (e.g. whether they focus on high achievers in keeping with the "model minority" stereotype or use the average among Asian students). Finally, including classrooms with many Asian students raises the concern of neighborhood- or school-specific trends, since a shift from zero to multiple Asian students may indicate local demographic shifts. Note that we include spillover results using classrooms with two or more Asian students in the Appendix for comparison.

tributes are predictive of Asian student exposure in the unadjusted model, the inclusion of teacher-school-grade-course fixed effects in Column 2 attenuates the magnitudes of coefficients and renders all coefficients insignificant with the exception of share economically disadvantaged (Appendix Table A1).

4 Results

4.1 Racial Differences in Teacher Assessments

Table 4 shows racial differences in teacher evaluations after adjusting for raw standardized test scores, individual characteristics, and class fixed effects. The outcome variable in the first column is an indicator for whether a teacher over-rates a student relative to their blind-scored achievement level, while the outcome variable in the second column is an indicator for whether a teacher under-rates a student.[16]

Results indicate teachers are 4.3 percentage points more likely to over-rate Asian students relative to White students in the same class with the same standardized test scores and individual characteristics. The magnitude is sizable, considering the effect is equivalent to nearly 12 percent of the baseline propensity of being over-rated of 0.347. We document comparable magnitudes when examining the phenomenon of under-rating. Teachers are 2.7 percentage points less likely to under-rate Asian students relative to White student counterparts who are observationally similar. This translates to a magnitude of 14 percent of the baseline propensity of being under-rated of 0.196.[17]

[16]Students who score a 4 are not included in the over-rating sample since it is mechanically infeasible to over-rate these students. Analogously, students who score a 1 are not included in the under-rating sample. The omitted racial group is White students.

[17]We also present alternative specifications that do not require conditioning the sample on the four-point achievement level scale, by using two different dependent variables. Table A2 replaces the over- and under-rating indicators with either the four-point teacher assessment scale or an indicator for teachers evaluating students at proficiency, defined as level 3 or above. Using the first outcome, we find that teachers confer Asian students a

Table 4: Racial Differentials in Teacher Assessments

	Over-rate $(NB > B)$	Under-rate $(NB < B)$
Baseline	0.347	0.196
Asian	0.043***	-0.027***
	(0.002)	(0.001)
Black	-0.021***	0.018***
	(0.001)	(0.001)
Hispanic	-0.020***	0.020***
	(0.001)	(0.001)
Native American	-0.021***	0.011***
	(0.003)	(0.002)
Other	-0.004***	0.004***
	(0.001)	(0.001)
N	11,830,325	13,539,719

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$. The omitted racial/ethnic student category is White students, so all coefficients can be interpreted relative to teachers' propensities to over-rate and under-rate White peers in columns (1) and (2), respectively. Teachers over-rate White students 34.1 percent of the time and under-rate White students 18.9 percent of the time.

In contrast to the favored ratings of Asian students, analogous racial

level of achievement that is 0.06 levels higher than same-scoring White classmates. This is a sizable difference given that Asian students' mastery as measured via EOG achievement levels is only 0.09 higher than that of White peers. Looking at the second outcome, this translates to about a 1.1 percentage point increase in the likelihood of being rated as proficient. Figure A2 provides visual evidence that this Asian-White assessment differential exists across a wide range of performance, although this figure does not control for student characteristics or classroom fixed effects.

differentials in teacher assessments go in the opposite direction for Black and Hispanic students, a finding that is consistent with previous literature on subjective teacher evaluations of under-represented minority students (Burgess & Greaves, 2013; Rangel & Shi, 2020). Notably, the magnitudes of teachers' increased propensity to over-rate and decreased propensity to under-rate Asians are at least as large, if not more so, than the extent of decreased over-rating and increased under-rating for Black and Hispanic students.[18]

Next, we disaggregate our results by math and reading classes to determine if racial differentials are driven by a particular subject. Appendix Table A4 shows that teachers are more likely to favor Asian students relative to observationally comparable White peers in both math and reading. Coefficients for reading are larger in magnitude. This may reflect the relatively more subjective nature of reading or English language arts instruction, which leaves more room for interpretation relative to the problem-based nature of mathematics. Separate analyses by grade level find similar patterns of teacher assessments in both elementary and middle schools (Appendix Table A5). Overall, these findings indicate that teachers' positive bias towards Asian students is pervasive across grades and subjects.[19]

[18]One potential concern for the interpretation of our results is presence of ceiling or floor effects, which could potentially lead End-of-Grade test scores not accurately characterizing student achievement at the top or bottom parts of the distribution. This in turn would suggest that comparisons in teacher ratings between students with the same raw End-of-Grade test scores at the top or bottom of the distribution may not actually be between students with comparable achievement levels. We assess raw test score distributions and find no evidence of significant floor or ceiling effects in standardized test scores in this sample, as evidenced by the fact that there appears to be neither censoring nor small numbers of discrete test scores that make a majority of observations at the tails (A. Ho & Yu, 2015). As a further robustness check, we re-run our analysis on the subset of students whose blind-scored achievement level scores are 2 or 3. Results of this analysis are shown in Table A3. Reassuringly, our findings are robust to this specification: for the sample of students who have blind score of level 2 or level 3, we find that teachers are qualitatively similar to our main estimations.

[19]We also looked to see whether our results are driven by either girls or boys. Results of this analysis are in Appendix Table A6 and indicate the presences of sizable disparities in teacher assessments of Asian students relative to White students for both girls and boys.

Robustness Checks

We undertake a number of analyses to address concerns that our results are consistent with alternative explanations. Specifically, we examine the roles of measurement error, differences in assessment standards across classes, unobserved behavioral characteristics, and racial biases in standardized testing that may potentially influence our results.

First, we address the issue that test scores may measure underlying ability with error and that findings on the Asian-White assessment gap may be partially attributable to this measurement error. This could be the case under the assumptions of racial differences in underlying skill distributions and uncorrelated errors (Hanushek & Rivkin, 2009). In this situation, White students who are observed as high-achieving will be more likely than observationally similar Asian students to be actually low-achieving. If teacher ratings reflect students' true achievement, teachers may be less likely to classify White students as high-achieving than Asian students, even in the absence of bias.

We examine the robustness of our results to measurement error concerns using an instrumental variables approach in Table B1. The first column replicates our main findings on racial differentials, while the second column shows that Asian-White gaps in teacher assessments are robust to including standardized test scores as a linear control interacted with the subject and grade level. Column 3 instruments for test scores using contemporaneous scores from the other subject, such as using same-year math scores to instrument for reading achievement. Under this specification, teachers are even more likely to favor Asians in over-rating and similarly likely to under-rate Asians compared to the OLS specification that does not correct for measurement error. One drawback of this first instrument is that it potentially suffers from an overly restrictive assumption of uncorrelated errors across contemporaneous subjects. For example, student illness and learning disruptions common to both subjects in a given year can contribute to correlated errors. Given these concerns, we next instrument using lagged achievement scores for the same subject. This specification enables accounting for measurement error under the assumption of uncorrelated errors over time and is not subject to concerns about contemporaneous shocks raised above. Results in Column 4 of Table B1 show coefficients that are very similar to the first instrument. Finally, we instrument for contemporaneous test scores using twice lagged scores in Column 5, which is perhaps even more likely to satisfy the assumption of uncorrelated errors. Once more, the likelihood that teachers over-rate Asian students relative to White students does not attenuate when taking measurement error into account. The Asian-White gap in under-rating is also robust across instrumental variables specifications. Taken together, the evidence suggests that measurement error in standardized testing does not explain our main findings.

Second, we address the concern that comparisons of blind and non-blind scores may be capturing differences in assessment standards across classes. Teachers' standards of skill mastery may vary depending on the particular school or classroom context, and this could generate racial gaps in teacher assessments in the presence of non-random sorting of students by race across schools and/or classrooms within schools. For instance, teachers with high-performing students may have higher standards for what constitutes a proficient student, independent of state guidelines. If this were the case, students in high-performing classes will be less likely to be over-rated than students in lower-performing classrooms with the same underlying ability, as measured by raw End-of-Grade test scores. While the inclusion of classroom fixed effects in our analyses control for differences across classroom in the outcome variables, they do not address classroom-level differences in baseline scores. To ensure that we are not mistaking these influences for teacher bias, we run an alternative specification in which our outcome variables, instead of being based on comparisons of teacher assessment of student achievement with the student's test-based assessment of achievement, compares teacher assessments with *adjusted* test-based assessments. These adjusted assessments are constructed by re-scaling test-based assessments of achievement levels within each class to match the distribution of teacher assessments of achievement levels (on the same four-point scale). Specifically, we use raw EOG scores to place the same number of students in each class into each blind-scored achievement level as observed in the corresponding teacher assessed achievement levels.[20] Table B2 shows that when we modify the outcomes of teacher over- and under-rating to use adjusted EOG achievement levels as an input, the estimated coefficients for Asian students are very similar to the unadjusted coefficients. This strongly suggests that what we interpret to be teacher bias is not confounded by differences in assessment standards across classes.

Third, we consider the possibility that systematic differences in teacher assessments of Asian and White students with the same standardized test score arise due to differences in unobserved characteristics, rather than teacher bias. To do so, we re-estimate our results using a variety of additional specifications in Table B4 in the Appendix. While our main specifications include detailed controls for students' prior disciplinary infractions, one concern could be that there are unobserved differences in severity of behavior within infraction categories, which could influence teachers' assessments of students. As a robustness check, we estimate a specification restricting to students with no prior infractions and find similar results to estimates on the full sample. Next, we augment the main specification with lagged test score controls to check whether entry-level achievement captures some other unobservable characteristic tied to achievement. We further augment our specification with lagged outcome measures in order to control for unobserved attributes influencing student achievement as well. We find results to be robust to this inclusion. Finally, we estimate a specification that controls for contemporaneous days absent and disciplinary

[20]For example, suppose a class has four students, and in the observed data, students have the following combinations of test-based assessments of achievement, teacher-based assessments of achievement, and raw EOG scores (B, NB, EOG): (4,4,112), (3,3,105), (2,4,78), (1,1,43). We re-scale the distribution of test-based assessments of achievement to match that of the distribution of teacher-based assessments using raw EOG scores, so the corresponding adjusted test-based assessments of achievement, teacher-based assessments of achievement, and raw EOG scores in this example would be (**adjusted-B**, NB, EOG): (**4**,4,112), (**4**,3,105), (**3**,4,78), (**1**,1,43).

infractions. These controls address the possibility that there may be unobserved shocks correlated with race that affect unobserved behavior and teacher assessments. One potential limitation of these controls is that contemporary behavior may be endogenous with teacher bias. Reassuringly, our results are robust to all of these specifications, providing further support for the validity of our findings. The fact that coefficients do not change much with these controls suggests that the directives that teachers receive to not take students' contemporaneous behavior into consideration in forming their assessments is working as intended.

Finally, we explore the possibility that our findings are driven by racial biases in standardized testing, rather than in teacher biases in evaluations. Theoretically, observed racial patterns in over-rating and under-rating are consistent with a scenario of standardized tests displaying negative cultural/racial bias towards Asian students in the absence of any teacher bias. If this were the case, we expect these results to be exacerbated for Asian students who do not speak English as their primary home language (relative to those who do speak English as a primary home language) for a couple of reasons. First, research indicates bilingual children may face especially large structural disadvantages with regards to standardized tests (Valdés & Figueroa, 1994). Additionally, home language can be seen as a proxy for assimilation, with the assumption that Asian students who speak English at home are less likely to suffer from cultural or Asian-specific racial biases that may be embedded in standardized tests. Our robustness check in Appendix Table B5 examines whether gaps are larger for Asian students who do not speak English as their primary home language. As with our main specifications, all estimates include controls for raw, uncoarsened standardized test scores. These controls address the possibility students that speak English at home are particularly skilled at reading or particularly skilled in other test-relevant dimensions. Furthermore, all estimations control for English home language, capturing baseline differences in bias between English and non-English speakers. Thus, the coefficient on Asian$\times$English captures the additional differential that teachers ascribe to Asian students who do not speak English at home, compared to White students who do not speak English at home. Results indicate that teachers are actually *more* likely to over-rate Asian students who report English as their primary home language and *less* likely to under-rate them. Our findings go in the opposite direction of the coefficients we would expect if results were being driven by racial bias in tests, providing further support that our findings reflect teacher bias.[21]

4.2 Heterogeneity in Teacher Assessments

Grouping Asian students into a single category potentially disguises their diverse experiences and trajectories. Existing studies examining the educational and labor market trajectories of Asians often rely on monolithic categories, even when research demonstrates substantial differences in schooling and earnings across Asian ethnic groups (Chiswick, 1983; Duleep & Sanders, 2012). In response, we take advantage of existing, albeit limited, data to investigate the extent to which teacher bias may vary across Asian ethnic groups. The NCERDC data do not contain direct information on a student's background beyond general racial and ethnic markers (White, Asian, Black, Hispanic, etc.), so we proxy for ethnic subgroups using two complementary methods. In the preferred specification, we rely on NCERDC data reporting a student's primary home language and use that information to classify Asian students into three regional subgroups: East Asian, South-

[21] One potential concern is that this interpretation of results does not take into consideration Asian students whose families come from countries where English is widely spoken and who might have unique cultural backgrounds despite speaking English at home. As a further check, we also run our analysis in Appendix Table B5 on a subset of counties in which the Asian population is least likely to be from Asian countries where English is widely spoken and find that our results are robust to this. We use detailed race information in ACS data from 2007-2013 to calculate what share of Asians in each county come from an Asian country that reports English as an official language, which includes India, Pakistan, Singapore, and the Philippines. Next, we re-run our specification of heterogeneity in teacher bias towards Asian students by English home language status using students from the subset of counties in which proportion of the Asian population that are from Asian countries where English is the official language is below the median. Estimates using this subsample are similar to those using the full sample.

east Asian, and South Asian.[22]

Table 5 shows the breakdown of Asian students in the sample by home language. Slightly over half of Asians in the sample report English as their primary language. Table 5 also provides descriptive statistics for Asian students by home language subgroup. Consistent with previously documented patterns, East Asian and South Asian students report a higher socioeconomic status than Southeast Asian students. They also have higher average math and reading scores.

Table 5: Asian Subgroups by Home Language Status

	N	Percent	% Econ Disadv.	Math scores	Reading scores
East Asian	4,153	10.70	0.22	1.10	0.46
South Asian	2,468	6.36	0.22	0.89	0.49
Southeast Asian	5,682	14.64	0.69	0.03	-0.33
Other Asian	2,299	5.93	0.67	-0.28	-0.59
Asian (English)	2,0726	53.42	0.30	0.72	0.46
Asian: Missing Language	3,471	8.95	0.45	0.53	0.21
Total/average	38,799	100.00	0.38	0.59	0.26

Observations denote unique students in grades 3-8 between 2007-2013 who identify as Asian. Classification by subgroup based on home language. For students who appear in the data for multiple years, we use the average economically disadvantaged status and average math and reading z-scores across years.

Next, we analyze teacher assessments across Asian subgroups using home language as a proxy for ethnicity. Table 6 shows substantial heterogeneity in the extent of teacher assessment gaps across subgroups. Compared to their assessments of White students, teachers are 6.1 percentage points more likely to over-rate South Asian students, 4.8 percentage more likely to over-rate East Asian students, and 2.4 percentage points more likely to over-rate Southeast Asian students. A Wald test indicates the coefficients between South Asians and Southeast Asians, and East Asians and Southeast Asians are significantly different at the 1 percent level, suggesting systematically lower prevalence of teachers over-rating Southeast Asian students

[22] Table C1 in the Appendix details the languages corresponding to each category. Most languages under the East Asian group are spoken in China, Japan, and South Korea. The majority of individuals in the South Asian group speak languages prevalent in India, Pakistan, and Bangladesh. The Southeast Asian group includes languages commonly spoken in Vietnam, Cambodia, Laos, Indonesia, Thailand, Malaysia, Philippines, and Burma.

relative to South Asian and East Asian peers, as proxied by home language. In terms of under-rating, estimates suggest that teachers are less likely to under-rate East, South, and Southeast Asians relative to White students. Coefficient estimates are not statistically different between any of the three groups, although the coefficient estimate is lowest in magnitude for Southeast Asian students, telling a qualitatively similar story to over-rating results.

Table 6: Differentials in Teacher Assessments by Home Language

	Over-rate $(NB > B)$	Under-rate $(NB < B)$
East Asian	0.048***	-0.023***
	(0.008)	(0.004)
South Asian	0.061***	-0.028***
	(0.006)	(0.004)
Southeast Asian	0.024***	-0.018***
	(0.003)	(0.002)
Other Asian	-0.014*	0.021***
	(0.008)	(0.008)
Asian: English	0.057***	-0.034***
	(0.003)	(0.001)
Asian: Missing language	0.042***	-0.033***
	(0.006)	(0.003)
N	11,830,325	13,539,719

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. Other minority races are included in regression, although they are not displayed in the table. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

A key advantage to using home language information to proxy for Asian ethnic subgroup is that we are able to infer detailed ethnic information at the individual level. However, a drawback of this approach is that a large portion of the sample reports English as their primary home language, and we are unable to infer detailed ethnic information for these students. We therefore analyze subgroup heterogeneity using a second approach based on Census ethnicity data. Specifically, we proxy for Asian subgroup concentration using the relative shares of East Asian, South Asian, and Southeast Asians in the county in which a school is located. This approach finds similar evidence of heterogeneity in teacher bias across Asian subgroups, with teachers being more positively biased towards South Asians and East Asians, relative to Southeast Asians. More details and results of this analysis can be found in Table C2 in Appendix C.

Additionally, we assess whether the degree of positive bias towards Asian students varies by school location. Table C3 in the Appendix augments our main specification with an interaction term for whether a school is based in a city (relative to a rural, town, or suburban location). Our findings reveal that teachers in cities are less positive towards Asian students: they are 1.4 percentage points less likely to over-rate Asian students and 0.9 percentage points more likely to under-rate Asian students than counterparts teaching in non-city settings. Upon closer examination, Table C4 shows that Asian students in cities have relatively lower socioeconomic and academic outcomes than White peers compared to Asian students outside of cities. This suggests that positive stereotyping towards Asian students may be lower in urban areas because Asians in these areas tend to conform less to the "model minority" stereotype, perhaps because of different compositions by ethnic subgroups.

Finally, we conduct analyses to examine the role of teacher characteristics. For example, teachers of a given racial and ethnic group or experience level may be more prone to classroom racial biases. We examine whether the extent of racial differentials is associated with teacher race, age, and experience and do not find any evidence that these attributes have significant bearing on teacher assessments towards Asian students. Results of this analysis are displayed in Appendix Table C5.[23] That there is no gradient between experience and tendencies to over-rate and under-rate by race suggests that these biases are pervasive, which is consistent with some ethnographic work in this area.[24]

4.3 Spillover Effects on Under-Represented Minorities

Despite the positive connotation of categorizing Asian students as a "model minority," such stereotypes may have adverse intrapersonal and interpersonal consequences, for example if teachers use these stereotypes to reinforce the notion of fundamental differences across groups and increase the usage of negative stereotypes (Kay et al., 2013). Table 7 investigates how exposure to Asian students affects teachers' assessments of students from *other* minority groups, relative to White peers with similar academic and behavioral records. Identification is based on variation in exposure to a single Asian student for a teacher in a given school who instructs a particular course (e.g., 5th grade math). Our models thus control for teacher attributes that are fixed at the teacher-school-grade-course level, including time-invariant preferences in assessments toward students of different racial and ethnic groups. The within-cell design addresses concerns involving non-random sorting of Asian students into classrooms on the basis of

[23] Note that due to the very small number of Asian teachers in our sample, we did not have enough statistical power to check for the role of racial congruence on our results. Such race match effects have been demonstrated in select contexts for Asian American students (see, for example: Lusher, Campbell, and Carrell (2018).

[24] For example, Drake (2022) interviewed teachers and students in an academically high-achieving secondary school about school culture and racialized expectations. One White teacher who has been instructing honors or AP-level courses for eight years had the following to say about a set of supplemental readings: "It's not like you need to read every word, okay? Relax; don't be all Asian about it." Another teacher who is Japanese American said the following about her Asian students, "I don't necessarily look at my classroom and treat a kid differently because they are Asian, but I know that if I have an Asian student in my classroom, I can count on that student. That student will probably work hard and be engaged. I can rely on that kid, and the parents, more so than I can for other [racial] groups." While this is admittedly a small sample, the research aligns with our empirical findings that differential teacher expectations by racial and ethnic groups are prevalent across White vs. non-White teachers as well as years of experience.

characteristics such as teacher race and course rigor.

To gauge the response of teachers to exposure to any Asian student, we restrict the analysis to classrooms with zero or one Asian student only. Table 7 shows that teachers respond to the presence of any Asian student in the classroom by decreasing their propensity to over-rate Black and Hispanic students relative to White students, compared to when no Asian students are present in the same teacher's classroom. Teachers are less likely to over-rate Black and Hispanic students by 0.5 and 0.6 percentage points, respectively. To place these magnitudes in context, teachers widen the Black-White and Black-Hispanic racial disparities in over-rating by approximately one-quarter (see Table 4) when they have an Asian student in their classroom. Teachers respond to the presence of an Asian student in the same classroom by increasing their propensity to under-rate by 0.5 percentage points among Black students. The relative change is on par with the magnitudes observed for over-rating. We do not find a significant corresponding change in under-rating for Hispanic students. These findings are especially troubling since teachers display sizable negative biases in assessments of Black and Hispanic students at baseline, suggesting these spillovers internalized by teachers are cumulatively disadvantageous for these underrepresented students.[25] To probe the channels underlying these effects, Table D2 examines whether teachers respond to exposure to an Asian student by changing pedagogy and classroom interactions in a manner that alters Black and Hispanic students' relative achievement, absenteeism, or disciplinary infractions. We find no indication that the presence of an Asian student affects these margins, suggesting that changes in teacher assessment gaps are not due to these channels.[26]

[25]Even though our main analyses restrict to classrooms with no more than one Asian student, Table D1 shows the results are robust to using the full sample.

[26]A caveat is that since our specification already accounts for students' contemporaneous achievement scores, the spillover effects we document are not inclusive of this important channel. Note that Column 1 in Table D2 shows statistically insignificant and negative coefficient estimates. The negative sign suggests that our estimates of spillover effects may even downplay the cumulative effect of teacher responses to having an Asian student in the classroom. Specifically, teachers may change their behaviors to widen the Black-White

Table 7: Effect of Exposure to One Asian Student

	Over-rate ($NB > B$)	Under-rate ($NB < B$)
Black×Any Asian	-0.005***	0.005***
	(0.002)	(0.002)
Hispanic×Any Asian	-0.006***	0.002
	(0.002)	(0.002)
Native American×Any Asian	-0.000	0.004
	(0.009)	(0.008)
Other×Any Asian	-0.004	-0.000
	(0.004)	(0.003)
Race×teacher-school-grade-course FE	Y	Y
N	10,614,152	11,789,383

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. Sample comprises students across all racial groups in grades 3-8 in odd-numbered columns and grades 4-8 in even-numbered columns between 2007-2013 in classrooms with zero or one Asian student. Any Asian is a binary variable indicating that the classroom had one Asian student. The omitted category is White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

A challenge to interpretation is that the negative spillover effects from Table 7 may derive from teacher exposure to individual attributes associated with being an Asian student, including academic achievement and family income, rather than exposure to Asian identity alone. The issue of cleanly separating out the influence of racial identity from the correlates of race is prevalent in the peer effects literature and not limited to this paper alone. We provide some indications of possible mechanisms, by first examining how spillover effects in teacher's assessments vary by the Asian student's academic achievement and socioeconomic status, before juxtaposing with the consequences of exposure to a Black or Hispanic student.

academic achievement gap in the classroom, in addition to exacerbating assessment gap after conditioning on test scores.

Table 8 analyzes whether teachers display negative spillovers in assessments of under-represented minority students in response to exposure to Asian students at particular parts of the achievement or income distribution. We use lagged test scores, normalized within the population of all students, as a measure for achievement to address potential endogeneity concerns with teacher expectations and Asian student performance.[27] Asian students are placed into four quartiles, from the lowest 25 percent (quartile 1) to the highest (quartile 4). Coefficients on the interactions between race variables and an Asian student of a particular quartile group are interpreted as differences in teacher propensities to over-rate or under-rate students in this racial group relative to White students when exposed to an Asian student in a given performance quartile, compared to when no Asian student is present.

The top panel in Table 8 shows that negative spillovers to teachers' propensity to over-rate Black and Hispanic students are concentrated among teachers exposed to the highest-achieving Asian students. Teachers respond to exposure to an Asian student in the top achievement quartile by decreasing their propensity to over-rate Black students by 2.0 percentage points relative to observationally similar White classmates. Teachers' response to this form of exposure nearly doubles the baseline estimated Black-White teacher over-rating gap. In contrast, teachers do not respond to the presence of an Asian student below the 75th percentile with changes in the propensity to over-rate Black students relative to White students. The analogous increase in the Hispanic-White over-rating gap from exposure to a high-performing Asian student is 1.6 percentage points, or four-fifth of the baseline difference in Table 4. Effects are more muted overall when looking at teacher under-rating. Teachers respond to the presence of an Asian student scoring above the 25th percentile by becoming 0.6 to 0.8 percentage points more likely to under-rate Black students relative to White peers. In contrast, there is no evidence of a significant change in the propensity for teachers to under-rate

[27] Sample sizes are smaller because we do not observe lagged scores for students in grade 3 and must restrict analyses to students in grades 4-8.

Table 8: Effect of Exposure to Asian Students, by Achievement and SES

	Over-rate $(NB > B)$	Under-rate $(NB < B)$
Panel A: By Achievement		
Black×Asian Lagged Z-score - Quartile 1	-0.002	-0.002
	(0.003)	(0.003)
Black×Asian Lagged Z-score - Quartile 2	0.003	0.008**
	(0.003)	(0.003)
Black×Asian Lagged Z-score - Quartile 3	-0.005	0.006**
	(0.003)	(0.003)
Black×Asian Lagged Z-score - Quartile 4	-0.020***	0.006*
	(0.004)	(0.003)
Hispanic×Asian Lagged Z-score - Quartile 1	-0.003	-0.002
	(0.004)	(0.004)
Hispanic×Asian Lagged Z-score - Quartile 2	0.001	0.005
	(0.004)	(0.004)
Hispanic×Asian Lagged Z-score - Quartile 3	-0.007	0.004
	(0.004)	(0.004)
Hispanic×Asian Lagged Z-score - Quartile 4	-0.016***	0.006
	(0.005)	(0.004)
N	8,798,325	9,786,625
Panel B: By Socioeconomic Status		
Black×High SES Asian	-0.007***	0.004**
	(0.002)	(0.002)
Black×Low SES Asian	-0.002	0.005**
	(0.002)	(0.002)
Hispanic×High SES Asian	-0.007**	0.005*
	(0.003)	(0.002)
Hispanic×Low SES Asian	-0.004	-0.000
	(0.003)	(0.003)
N	10,611,693	11,786,172
Race×Teacher-school-grade-course FE	Y	Y

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. Sample comprises students across all racial groups in grades 3-8 between 2007-2013 in classrooms with zero or one Asian student. Any Asian is a binary variable indicating that the classroom had one Asian student. Lagged Z-score is the Asian student's standardized lagged z-score normalized within the population of all students in a given grade and year. Quartiles are defined using the statewide sample of students at the year, grade, and subject level (e.g. quartile 1 is the 25th percentile or below). Low (high) SES assumes a value of 1 if the Asian student is (not) economically disadvantaged, defined as being eligible for free or reduced price lunch. The omitted category is White students. Models include Native American and students of other racial or ethnic groups and their interactions with the Any Asian, Asian Lagged Z-score quartiles, and Low SES Asian variables. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage.

Hispanic students as a result of having an Asian student enrolled in class. The bottom panel of Table 8 examines teacher exposure to high- vs. low-SES Asian students, defined based on economic disadvantage status. We find that negative spillover effects in teacher assessments are concentrated among exposure to economically advantaged Asian students.

A lingering question is whether the negative spillover effects are uniquely driven by the presence of Asian students or whether these spillover effects will exist as long as teachers are exposed to the correlates of this racial group, namely higher academic achievement and family income. To assess this, we examine whether the presence of a single Black or Hispanic student leads to similar spillover effects on teacher assessments. Tables D3 and D4 in the Appendix restrict the sample to classes with zero or a single Black student to assess how teachers' ratings of Hispanic students change in response to an exposure to a Black student in the classroom, as well as how these effects vary by the academic performance and socioeconomic status of the Black student. The presence of a Black student in a classroom has no effect on the propensity of teachers to over-rate or under-rate Hispanic students, a result that is distinct from the racial disparity-exacerbating effects of exposure to an Asian student (Table D3). Table D4 shows null effects on teacher assessments across both the achievement and SES gradients for the given Black student. There are similarly no spillover effects to teachers' assessments of Black students relative to White students from exposure to a Hispanic student (Table D5). Further investigation in Table D6 shows that exposure to a Hispanic student at the lowest achievement quartile bridges the Black-White over-rating gap, while teachers respond to having a high-achieving Hispanic student by exacerbating the disparity. This symmetry stands in contrast to spillover effects concentrated among teachers' exposure to high-performing Asian students.

Taken together, these results suggest that teachers' responses to exposure to an Asian student are distinct from exposure to other, under-represented, minority groups. The existence of significant negative spillovers sets perceptions of Asian students apart. Notably, these effects are driven by teacher reactions to the presence of Asian students in the classroom whose high achievement conforms to the "model minority" characterization, rather than to students who defy the stereotype. Such findings underscore the challenges of separately identifying mechanisms. These effects are likely not driven entirely by responses to individual attributes such as achievement, since we do not observe analogous negative spillovers from exposure to high-achieving Black or Hispanic students. Teachers' reactions to Asian identity, which is inextricably linked to assumptions about achievement, is likely a necessary component for the effects we observe. As such, we interpret these results as the cumulative impact of classroom exposure to Asian identity and associated student attributes ranging from performance and behaviors to family background.

5 Conclusion

Limited research exists on Asian Americans, despite their increasing prominence in K-12 education and status as the single fastest growing racial demographic group in the United States. This study provides evidence for the treatment of Asian Americans as "model minorities" in elementary and secondary schools. We show that teachers, when tasked with assessing student mastery in a subject, rate Asian students more favorably relative to White students in the same class with the same standardized test scores. The assessment advantages conferred upon Asian students are persistent across grade levels and subjects and are robust to accounting for factors such as measurement error and behavioral differences. Crucially, teacher assessment patterns that set Asians apart from other groups of minority students can have lasting consequences given the influence of teacher expectations on how teachers treat students, students' own behaviors, and subsequent longer-term academic trajectories (Botelho, Madeira, & Rangel, 2015; Card & Giuliano, 2016; Hill & Jones, 2021; Lindahl, 2016; Papageorge et al., 2020).

We investigate potential consequences of this so-called positive bias by examining the extent to which teacher assessments of Asian students might interact with their judgment of students belonging to other minority groups. We find that teachers respond to exposure to an Asian student by depressing their assessments of Black and Hispanic students relative to White counterparts, leading to a widening in both Black-White and Hispanic-White assessment gaps. This suggests that there are significant negative consequences of teachers demonstrating positive bias towards Asian students. These findings recall small-scale studies demonstrating that positive stereotypes reinforce beliefs in the biological underpinnings of group differences and the application of negative stereotypes (Kay et al., 2013) and suggest the potential for negative spillover effects of biases with an ostensibly positive connotation. To the extent that stereotypes are based on representative generalizations that are exaggerated to provide the greatest differentiation in a given context (Bordalo et al., 2016; Kahneman & Tversky, 1972), teachers' stereotypical judgment for Black and Hispanic students may be most salient when they encounter a high-performing Asian student.

Taken together, our results underscore the existence and potential pitfalls of positive biases. Future work can explore the long-term consequences of positive biases for Asian students themselves, building on previous research that establish substantial intrapersonal and interpersonal costs of receiving positive stereotypes.[28] Despite theory and evidence from mostly lab settings that positively stereotyped group members may change their academic expectations and orientation towards particular academic or career tracks (Czopp, 2010; C. Ho et al., 1998), little research links these short-term changes in expectations and behaviors to long-run academic outcomes. A related topic that merits additional research is the extent of differential responses among individuals who conform in varying degrees to positive stereotypes of the larger group; namely, shifting away from a monolithic conception of Asian students to distinguish between the academic responses of Asian subgroups. Finally, future work that focuses on understanding the

[28]Previous studies have shown that the targets of such biases are more likely to experience psychological distress and depersonalization and are less likely to seek help from others (e.g. Gupta, Szymanski, and Leong (2011)).

inception of the model minority stereotype and factors that influence teachers' biases will have important implications for policy.

References

Altonji, J. G., & Blank, R. M. (1999). *Race and gender in the labor market* (Handbook of Labor Economics). Elsevier.

Arcidiacono, P., Kinsler, J., & Ransom, T. (2020). Asian American Discrimination in Harvard Admission. *Working Paper*.

Bifulco, R., Fletcher, J. M., & Ross, S. L. (2011). The Effect of Classmate Characteristics on Post-secondary Outcomes: Evidence from the Add Health. *American Economic Journal: Economic Policy*, *3*(1), 25–53.

Black, D. A., Haviland, A. M., Sanders, S. G., & Taylor, L. J. (2008, July). Gender Wage Disparities among the Highly Educated. *Journal of Human Resources*, *43*(3), 630–659.

Black, S., Devereux, P., & Salvanes, K. G. (2013). Under Pressure? The Effect of Peers on Outcomes of Young Adults. *Journal of Labor Economics*, *31*(1), 119–153.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016, November). Stereotypes. *The Quarterly Journal of Economics*, *131*(4), 1753–1794.

Botelho, F., Madeira, R. A., & Rangel, M. A. (2015). Racial Discrimination in Grading: Evidence from Brazil. *American Economic Journal: Applied Economics*, *7*(4), 37–52.

Budiman, A. (2020). *Asian Americans are the fastest-growing racial or ethnic group in the U.S. electorate* [Pew Research Center].

Burgess, S., & Greaves, E. (2013). Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities. *Journal of Labor Economics*, *31*(3), 535–576.

Card, D., & Giuliano, L. (2016). Can Tracking Raise the Test Scores of High-Ability Minority Students? *American Economic Review*, *106*(10), 2783–2816.

Chang, D., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology*, *71*(2), 235–242.

Cheryan, S., & Bodenhausen, G. V. (2016). When Positive Stereotypes

Threaten Intellectual Performance: The Psychological Hazards of "Model Minority" Status:. *Psychological Science*.

Chiswick, B. R. (1983). An Analysis of the Earnings and Employment of Asian-American Men. *Journal of Labor Economics*, *1*(2), 197–214.

Czopp, A. M. (2010). Studying is lame when he got game: racial stereotypes and the discouragement of Black student-athletes from schoolwork. *Social Psychology of Education*, *13*(4), 485–498.

Donovan, S., & Cross, C. (2002). *Minority Students in Special and Gifted Education*. National Academy Press.

Drake, S. (2022). *Academic apartheid: Race and the criminalization of failure in an american suburb*. Oakland CA: Russell Sage Foundation.

Duleep, H. O., & Sanders, S. (1992). Discrimination at the Top: American-Born Asian and White Men. *Industrial Relations*, *31*(3), 416–432.

Duleep, H. O., & Sanders, S. G. (2012, June). *The Economic Status of Asian Americans Before and After the Civil Rights Act* (SSRN Scholarly Paper No. ID 2089668). Rochester, NY: Social Science Research Network.

Fejgin, N. (1995). Factors Contributing to the Academic Excellence of American Jewish and Asian Students. *Sociology of Education*, *68*(1), 18–30.

Francis, D. (2012). Sugar and spice and everything nice? teacher perceptions of black girls in the classroom. *Review of Black Political Economy*, *39*, 311–320.

Francis, D., de Oliveira, A., & Dimmitt, C. (2019). Do school counselors exhibit bias in recommending students for advanced coursework? *The B.E. Journal of Economic Analysis & Policy*, *19*(4), 1–17.

Gupta, A., Szymanski, D. M., & Leong, F. T. L. (2011). The "model minority myth": Internalized racialism of positive stereotypes as correlates of psychological distress, and attitudes toward help-seeking. *Asian American Journal of Psychology*, *2*(2), 101–114.

Hanushek, E. A., & Rivkin, S. G. (2009, March). Harming the best: How schools affect the black-white achievement gap. *Journal of Policy Analysis and Management*, *28*(3), 366–393.

Hilger, N. (2017). Upward Mobility and Discrimination: the Case of Asian

Americans. *NBER Working Paper 22748*.

Hill, A. J., & Jones, D. B. (2021, June). Self-Fulfilling Prophecies in the Classroom. *Journal of Human Capital*, 000–000.

Ho, A., & Yu, C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365–388.

Ho, C., Driscoll, D., & Loosbrock, D. (1998). Great Expectations: The Negative Consequences of Falling Short1. *Journal of Applied Social Psychology*, *28*(19), 1743–1759.

Hsin, A., & Xie, Y. (2014). Explaining Asian Americans' academic advantage over whites. *Proceedings of the Natironal Acadmy of Sciences of the United Staets of America*, *111*(23), 8416–8421.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *9*(2), 131–155.

Jussim, L., & Harber, K. D. (2016). Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies:. *Personality and Social Psychology Review*.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Kao, G. (1995). Asian Americans as Model Minorities? A Look at Their Academic Performance. *American Journal of Education*, *103*(2), 121–159.

Kay, A. C., Day, M. V., Zanna, M. P., & Nussbaum, A. D. (2013). The insidious (and ironic) effects of positive stereotypes. *Journal of Experimental Social Psychology*, *49*(2), 287–291.

Lavy, V. (2008, October). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, *92*(10–11), 2083–2105.

Lavy, V., & Megalokonomou, R. (2019, June). *Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study* (Tech. Rep. No. w26021). Cambridge, MA:

National Bureau of Economic Research.
Lavy, V., & Sand, E. (2018, November). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, *167*, 263–279.
Lee, J., & Zhou, M. (2015). *The Asian American achievement paradox*. New York: Russell Sage Foundation.
Lindahl, E. (2016). Are teacher assessments biased? – evidence from Sweden. *Education Economics*, *24*(2), 224–238.
Lusher, L., Campbell, D., & Carrell, S. (2018, March). TAs like me: Racial interactions between graduate teaching assistants and undergraduates. *Journal of Public Economics*, *159*, 203–224.
Mar, D. (2005). Asian Americans in the Labor Market: Public Policy Issues. *AAPI Nexus Journal*, *3*(2).
Ouazad, A. (2014, May). Assessed by a Teacher Like Me: Race and Teacher Assessments. *Education Finance and Policy*, *9*(3), 334–372.
Ouazad, A., & Page, L. (2013, September). Students' perceptions of teacher biases: Experimental economics in schools. *Journal of Public Economics*, *105*, 116–130.
Papageorge, N., Gershenson, S., & Kang, K. M. (2020). Teacher Expectations Matter. *Review of Economics and Statistics*, *102*(2), 234–251.
Rangel, M., & Shi, Y. (2020). First Impressions: The Case of Teacher Racial Bias. *Working Paper*.
Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the Classroom. *The Urban Review*, *3*(1), 16–20.
Sakamoto, A., Goyette, K. A., & Kim, C. H. (2009). Socioeconomic Attainments of Asian Americans. *Annual Review of Sociology*, *35*, 255–276.
Sakamoto, A., Woo, H., & Yap, K.-L. (2006). Are Native-born Asian Americans Less Likely To Be Managers? Further Evidence on the Glass-ceiling Hypothesis – AAPI Nexus Journal. *AAPI Nexus*, *4*(1), 13–37.
Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Westport, CT, US: Ablex Publishing.
Weinberger, C. J. (1998, January). Race and Gender Wage Gaps in the Market

for Recent College Graduates. *Industrial Relations: A Journal of Economy and Society*, *37*(1), 67–84.

Wu, E. (2014). *The Color of Success: Asian Americans and the Origins of the Model Minority*. Princeton University Press.

Xie, Y., & Goyette, K. A. (2004). *A Demographic Portrait of Asian Americans*. Russell Sage Foundation.

APPENDIX

A Additional Figures and Tables

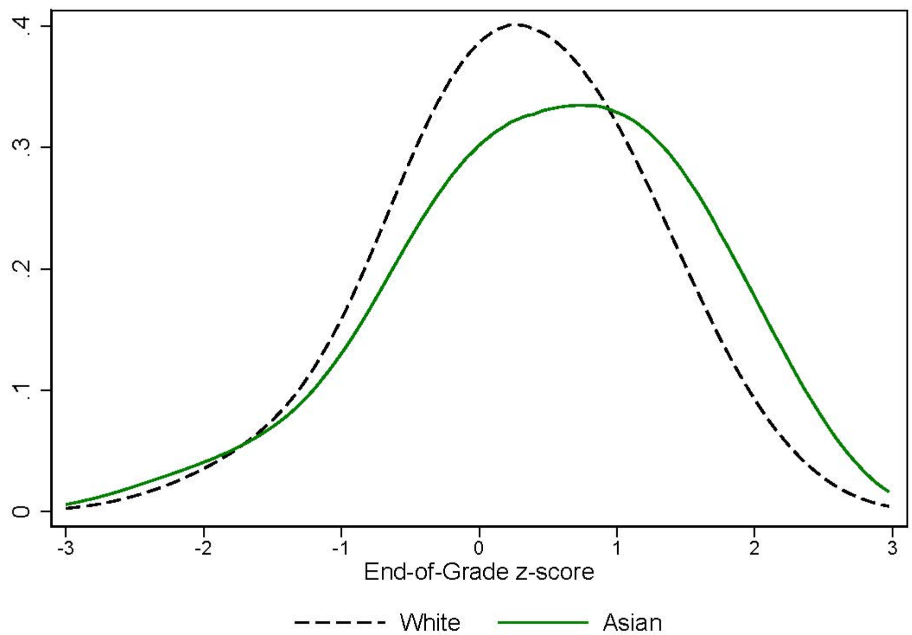


Figure A1: Performance Distribution by Race

Note: Sample comprises students in grades 3-8 between 2007-2013. End-of-Grade score density curves are estimated for White and Asian students using kernel density estimation.

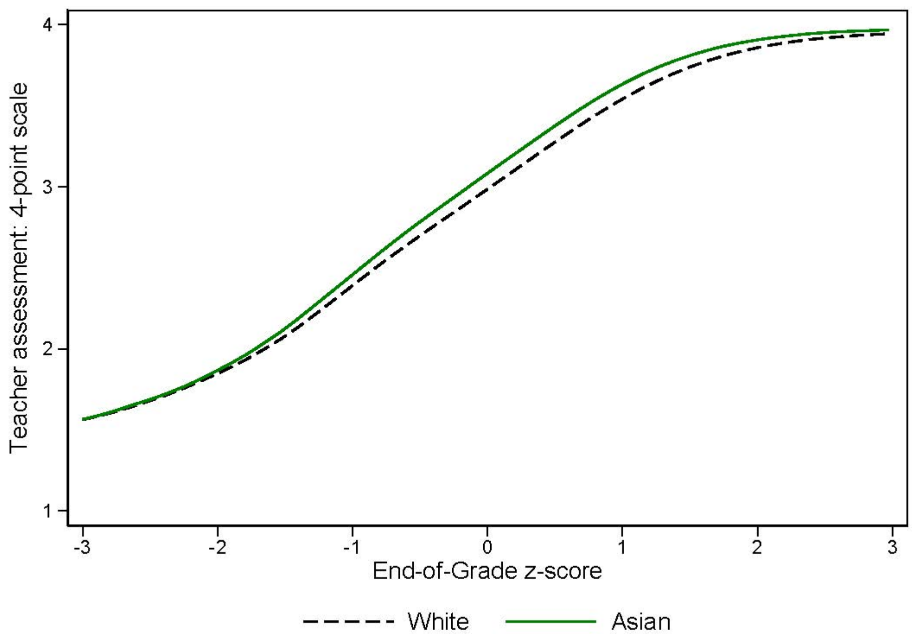


Figure A2: Teacher Assessments and Standardized Test Scores by Race

Note: Sample comprises students in grades 3-8 between 2007-2013. Bivariate relationship between teacher assessments on the four-point scale and raw End-of-Grade test scores estimated via local polynomial regressions.

Table A1: Variation in Exposure to Any Asian Student

	Any Asian Student	
	(1)	(2)
Share Female	0.006	0.008
	(0.010)	(0.012)
Share White	0.628***	-0.009
	(0.153)	(0.220)
Share Black	0.809***	-0.128
	(0.153)	(0.221)
Share Hispanic	0.744***	-0.075
	(0.153)	(0.221)
Share Native American	0.476***	-0.116
	(0.154)	(0.226)
Share Other	1.115***	-0.044
	(0.156)	(0.221)
Share Econ. Disadvantaged	-0.240***	-0.054***
	(0.008)	(0.010)
White-Black Achievement Gap	0.009***	0.004
	(0.002)	(0.003)
White-Hispanic Achievement Gap	0.005**	0.002
	(0.002)	(0.002)
Teacher-school-grade-course FE	N	Y
N	320,448	309,453

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. Classroom sample includes grades 4-8 from 2007-2013, and is limited to classrooms that have either zero or one Asian student. We exclude Asian students from the shares of students by race and ethnicity to avoid mechanically-induced changes based on exposure to an Asian student. Achievement gaps are computed as the difference in the average lagged z-scores across racial or ethnic groups.

Table A2: Racial Differentials in Teacher Assessments: Alternative Dependent Variables

	Teacher Assessment (Four-point Scale)				Proficiency (Level 3+)	
	(1)	(2)	(3)	(4)	(5)	(6)
Asian	0.075***	0.063***	0.056***	0.060***	0.011***	0.013***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Black	-0.051***	-0.058***	-0.040***	-0.040***	-0.016***	-0.017***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Hispanic	-0.061***	-0.061***	-0.041***	-0.042***	-0.020***	-0.021***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Native American	0.004	-0.051***	-0.031***	-0.032***	-0.008***	-0.009***
	(0.008)	(0.004)	(0.004)	(0.004)	(0.002)	(0.003)
Other	-0.021***	-0.023***	-0.007***	-0.007***	-0.001	-0.001
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
N	15,988,137	15,968,445	15,232,063	15,232,088	15,232,063	15,232,088
Raw End-of-Grade test score FE	Y	Y	Y		Y	
Class FE		Y	Y	Y	Y	Y
Individual characteristics			Y	Y	Y	Y
Quartic polynomial - EOG z-scores				Y		Y

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. Individual characteristics include gender, an indicator for economic disadvantage, lagged days absent, and lagged disciplinary infractions.

Table A3: Racial Differentials in Teacher Assessments: Students in Achievement Levels 2 and 3 Only

	Over-rate ($NB > B$)	Under-rate ($NB < B$)
Asian	0.051***	-0.016***
	(0.002)	(0.001)
Black	-0.023***	0.012***
	(0.001)	(0.001)
Hispanic	-0.021***	0.015***
	(0.001)	(0.001)
Native American	-0.023***	0.004*
	(0.003)	(0.002)
Other race	-0.005***	0.001
	(0.001)	(0.001)
N	10,137,865	10,137,865

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include controls for gender, economic disadvantage, lagged days absent, and lagged disciplinary infractions. The sample of students in the assessment of teacher over-rating and under-rating include those with $B \in \{2, 3\}$.

Table A4: Racial Differentials in Teacher Assessments, by Subject

	Math		Reading	
	Over-rate $(NB > B)$	Under-rate $(NB < B)$	Over-rate $(NB > B)$	Under-rate $(NB < B)$
Asian	0.030***	-0.022***	0.049***	-0.031***
	(0.002)	(0.002)	(0.002)	(0.002)
Black	-0.010***	0.009***	-0.030***	0.026***
	(0.001)	(0.001)	(0.001)	(0.001)
Hispanic	-0.016***	0.020***	-0.024***	0.019***
	(0.001)	(0.001)	(0.001)	(0.001)
Native American	-0.014***	0.007***	-0.027***	0.013***
	(0.003)	(0.003)	(0.004)	(0.003)
Other	0.000	-0.000	-0.007***	0.007***
	(0.001)	(0.001)	(0.002)	(0.001)
N	5,105,363	6,173,016	6,724,949	7,366,689

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

Table A5: Racial Differentials in Teacher Assessments, by Grade Level

	Elementary		Middle	
	Over-rate $(NB > B)$	Under-rate $(NB < B)$	Over-rate $(NB > B)$	Under-rate $(NB < B)$
Asian	0.037***	-0.025***	0.048***	-0.029***
	(0.003)	(0.002)	(0.002)	(0.001)
Black	-0.023***	0.020***	-0.020***	0.017***
	(0.001)	(0.001)	(0.001)	(0.001)
Hispanic	-0.015***	0.019***	-0.025***	0.020***
	(0.001)	(0.001)	(0.001)	(0.001)
Native American	-0.022***	0.006*	-0.021***	0.016***
	(0.005)	(0.003)	(0.003)	(0.002)
Other	-0.001	0.004**	-0.006***	0.004***
	(0.002)	(0.002)	(0.001)	(0.001)
N	5,580,854	6,374,394	6,249,216	7,165,083

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students enrolled in an elementary school (grades 3-5) or middle school (grades 6-8) in 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

Table A6: Racial Differences in Teacher Assessments: Heterogeneity by Gender

	Boys		Girls	
	Over-rate $(NB > B)$	Under-rate $(NB < B)$	Over-rate $(NB > B)$	Under-rate $(NB < B)$
	(1)	(2)	(3)	(4)
Asian	0.038***	-0.029***	0.047***	-0.027***
	(0.003)	(0.002)	(0.003)	(0.002)
Black	-0.023***	0.022***	-0.019***	0.015***
	(0.001)	(0.001)	(0.001)	(0.001)
Hispanic	-0.024***	0.023***	-0.018***	0.016***
	(0.001)	(0.001)	(0.001)	(0.001)
Native American	-0.016***	0.015***	-0.025***	0.007**
	(0.004)	(0.003)	(0.004)	(0.003)
Other race	-0.002	0.005***	-0.006***	0.002
	(0.002)	(0.001)	(0.002)	(0.001)
N	5,984,262	6,714,646	5,796,960	6,792,198

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, days absent, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

B Robustness Checks

Table B1: Role of Measurement Error

			Instrumental variables		
			Other	Lagged	Twice Lagged
	OLS	OLS	Subject	Subject	Subject
	(1)	(2)	(3)	(4)	(5)
Over-rate: $(NB > B)$					
Asian	0.043***	0.038***	0.060***	0.058***	0.058***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	11,830,325	11,830,110	11,768,348	9,769,914	7,488,893
Under-rate: $(NB < B)$					
Asian	-0.027***	-0.029***	-0.026***	-0.026***	-0.027***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	13,539,719	13,539,719	13,496,699	11,322,103	8,669,605
Raw End-of-Grade score FE	Y				
EOG z-score		Y	Y	Y	Y

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics and class fixed effects. Student characteristics include gender, an indicator for economic disadvantage, lagged days absent, and lagged disciplinary infractions. The first column controls for raw end-of-grade test score fixed effects interacted with subject, grade, and year. Column 2 includes EOG z-scores at the subject-grade-year level, entered linearly. Column 3 instruments for EOG z-scores using the contemporaneous other subject z-score (i.e., instrument current math z-scores using current reading z-scores). Column 4 instruments for EOG z-scores using lagged same-subject z-scores, while Column 5 instruments for z-scores using twice lagged same-subject z-scores. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

Table B2: Adjusted Blind Achievement Levels

	(1) Over-rate $(NB > B)$	(2) Under-rate $(NB < B)$
Asian	0.040***	-0.027***
	(0.002)	(0.001)
Black	-0.019***	0.016***
	(0.001)	(0.001)
Hispanic	-0.023***	0.018***
	(0.001)	(0.001)
Native American	-0.014***	0.014***
	(0.003)	(0.002)
Other	-0.003**	0.003***
	(0.001)	(0.001)
N	11,295,848	14,319,841

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. Specification uses raw EOG test scores to put students into adjusted achievement levels such that the number of students per class in each level is the same as the number of students at each of the four teacher rating levels. Outcomes are indicator variables for whether the teacher rating level is higher or lower than the *adjusted* blind-scored achievement levels based on EOG performance. The sample of students in the assessment of teacher over-rating includes those with $Adjusted\ B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $Adjusted\ B \in \{2, 3, 4\}$.

Table B3: Disciplinary Infractions List

Infraction	Frequency
Disruptive behavior	1,693,620
Bus misbehavior	806,673
Insubordination	643,970
Aggressive behavior	642,685
Fighting	582,034
Inappropriate language/disrespect	537,929
Disrespect of faculty/staff	435,807
Other school defined offense	253,873
Other	169,684
Bullying	132,511
Theft	119,418
Excessive tardiness	101,421
Disorderly conduct	80,255
Dress code violation	78,637
Skipping class	71,356
Late to class	62,470
Cell phone use	62,076
Communicating threats	61,960
Skipping school	60,386
Inappropriate items on school property	54,307
Assault on student	50,019
Property damage	48,119
Harassment–verbal	47,428
Harassment–sexual	39,740
Possession of a weapon (excluding firearms/explosives)	36,941
Honor code violation	31,200
Truancy	25,818
Being in an unauthorized area	22,959
Leaving school without permission	20,634
Excessive display of affection	18,708
Falsification of information	18,333
Leaving class without permission	18,169
Unlawfully setting a fire	17,469
Assault on student w/o weapon and not resulting in injury	17,290
Misuse of school technology	17,095
Gang activity	12,167
Possession of tobacco	10,437
Possession of controlled substance–marijuana	9,872
Affray	8,561
Cutting class	7,844
Immunization	7,800
Repeat Offender	7,115
Assault–other	6,356
Assault on school personnel not resulting in injury	6,057
Possession of counterfeit items	5,729
Use of tobacco	5,408
Mutual sexual contact between two students	3,562
Alcohol possession	3,082
Hazing	2,805
Possession of controlled substance–other	2,717

Table displays list of disciplinary infractions that students can be reported for, as well as the frequency with which each infraction appears in the sample. A given student may have been reported for multiple types of infractions over the course of the year, and it is also possible for a student to be reported for the same infraction multiple times over the course of the year. Note: we restrict this list to the 50 most frequently occurring infraction types in the data.

Table B4: Racial Differentials in Teacher Assessments: Alternative Specifications

	Full Sample	No Prior Infractions	Full Sample		
	(1)	(2)	(3)	(4)	(5)
Over-rate: $(NB > B)$					
Asian	0.043***	0.042***	0.064***	0.063***	0.036***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	11,830,325	10,178,351	9,769,850	9,757,655	11,830,084
Under-rate: $(NB < B)$					
Asian	-0.027***	-0.027***	-0.036***	-0.035***	-0.023***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	13,539,719	12,067,816	11,322,074	11,311,774	13,539,471
Test score FE × subject × grade × year	Y	Y	Y	Y	Y
Lagged test score FE × subject × grade			Y	Y	
Lagged teacher judgment				Y	
Contemporaneous Absence+Infractions					Y

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, an indicator for economic disadvantage, lagged days absent, and lagged disciplinary infractions. Column 2 restricts the sample to students with no infractions in the prior year. Column 3 augments the main sample with lagged end-of-grade test score fixed effects interacted with subject, grade, and year. The sample size drops because we do not observe lagged test scores for some students in the sample, in particular students in third grade. Estimation results in Column 1 are very similar when we re-estimate the specification on the sample of students for whom we observe lagged test scores. Column 4 further augments column 3 with lagged teacher over-rating controls in the top panel and lagged teacher under-rating controls in the bottom panel. Column 5 augments the main sample with contemporaneous controls in days absent and number of disciplinary infractions. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

Table B5: Restrict to Students who Report English Home Language

	Full Sample		Restricted Sample	
	Over-rate	Under-rate	Over-rate	Under-rate
	$(NB > B)$	$(NB < B)$	$(NB > B)$	$(NB < B)$
	(1)	(2)	(3)	(4)
Asian	0.035***	-0.024***	0.029***	-0.013***
	(0.003)	(0.002)	(0.005)	(0.004)
Asian×English	0.020***	-0.009***	0.021***	-0.017***
	(0.004)	(0.002)	(0.007)	(0.005)
English	0.008***	-0.006***	0.012***	-0.009***
	(0.001)	(0.001)	(0.002)	(0.002)
N	11,830,108	13,539,488	3,424,435	3,826,373

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. Other minority groups and their interactions with English home language are included in the regression, although they are not displayed in the table. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$. In the estimations for the restricted sample in columns (3) and (4), we only include the students in the subset of counties in which proportion of the Asian popuation that are from Asian countries where English is the official language is below the median, calculated using 2007-2013 ACS data.

C Heterogeneity in Teacher Assessments

Table C1 shows how NCERDC self-reported primary home languages are categorized into the ethnic subgroups of East Asian, South Asian, and Southeast Asian. In addition to the home languages in the table, some Asian students in this sample also reported English as a primary home language or a non-English language that was not identifiable as a language associated with an Asian ethnic subgroup (e.g., Italian or Swahili). Table 5 in the paper shows the breakdown of Asian students in the sample by reported language categories.

Table C1: NCERDC Home Language Code Classification

Subroup	**Language Codes**
East Asian	Chinese (Mandarin), Chinese (Cantonese), Chinese (Zhongwen), Chinese (Shanghai/Wu), Chinese (Taiwan), Chinese, Japanese, Korean
South Asian	Gujarati, Hindi, Punjabi/Panjabi, Tamil, Telugu, Urdu, Bengali, Bihari, Hindi/Indian/Urdu, Kannada, Kashmiri, Pushto/Eastern Pashto, Saurashtra/Sowrashtra, Sindhi, Marathi, Oriya, Hindko
Southeast Asian	Vietnamese, Burmese, Cambodian/Khmer, Cebuano, Indonesian, Hmong/Hmong-Mien/Hmogie/Chaug, Koho, Rade, Tagalog/Filipino, Lahu, Lao/Laotian, Tai/Eastern Tai, Malay/Bahasa Malaysia, Malayalam, Thai/Ta/Thaiklang, Jarai, Mnong, Chin

Classification of Asian students into subgroups based on NCERDC self-reported home language.

As an alternative approach, we use county-level Asian subgroup population to proxy for students' ethnicities. Data comes from the American Community Survey (ACS) from 2007-2013. For each county, we measure the average aggregate Asian population over that time frame, as well as the Asian population broken down by subgroup (East Asian, South Asian, and Southeast Asian). We use the proportion of Asians of a given subgroup in the county as a proxy for how likely an Asian student is from a given subgroup. One limitation of this approach is that the data are rather coarse —unlike in our preferred approach, we do not observe ethnicity data at the individual level. Furthermore, the ACS only has individual county-level data for the 25 largest counties in North Carolina, out of 50 total. The remaining smaller counties are aggregated into one category. The benefit of this approach though, is that we are able to circumvent the issue that many Asians in our sample are English-speaking, which is a shortcoming in the home language approach.

Table C2 shows results using county-level Asian ethnic shares as a subgroup proxy. Results indicate that a 10 percentage point increase in the share of Asians in a county that are East Asian, relative to Southeast Asian, increases the propensity that a teacher will over-rate an Asian student by 0.5 percentage points. Furthermore, a 10 percentage point increase in the share of Asians in a county that are South Asian, relative to Southeast Asian, decreases the propensity that a teacher will under-rate an Asian student by 0.5 percentage points. Estimated coefficients on the effects of South Asians on teacher ove-rating and East Asians on teacher under-rating and small and insignificant. These findings are consistent with results that home language as a proxy in suggesting that teachers are more positive in their assessments of South and East Asians relative to Southeast Asians.

Conversely, a 10 percentage point increase in the share of Asians in a county that are South Asian, relative to Southeast Asian, decreases the propensity that a teacher will under-rate an Asian student by 0.6 percentage points. We find no statistically significant effect of an increase in East Asian share on the propensity that a teacher under-rates a Southeast Asian student. A Wald test of coefficients shows that the effect of proportion East Asian and proportion South Asian are not statistically different from one another at the 5% level but are different at the 10% level.

Next, Table C3 examines whether racial gaps in teacher assessment differ for teachers in an urban versus more rural setting. To do so, we augment our main specification with an interaction term for whether the school is based in a city (relative to a rural, town, or suburban location), as classified by NCERDC. Results indicate teachers in cities are less positive towards Asian students: they are 1.4 percentage points less likely to over-rate Asian students and 0.9 percentage points more likely to under-rate Asian students than counterparts teaching in non-city settings. Teachers in cities are also less positive towards Black and Hispanic students.

Table C2: Racial Differentials in Teacher Assessments by ACS Asian Subgroup

	Over-rate $(NB > B)$	Under-rate $(NB < B)$
Asian	0.040***	-0.023***
	(0.008)	(0.005)
Asian×Proportion Asian	-0.007	0.029***
	(0.012)	(0.007)
Asian×Proportion East|Asian	0.005**	-0.001
	(0.002)	(0.001)
Asian×Proportion South|Asian	-0.000	-0.005***
	(0.002)	(0.001)
Class FE	Y	Y
Race×teacher FE	Y	Y
N	12,383,463	14,147,869

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. Other minority races and interactions with Asian share and Asian subgroup shares are included in regression, although they are not displayed in table. Coefficients represent the effect of a 10 percentage point increase in proportion of interest. The omitted Asian subgroup share is proportion of Southeast Asians. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. Asian subgroups are classified using reported ancestry data from the ACS (East Asian: Chinese, Cantonese, Japanese, Okinawan, Korean, Taiwanese. South Asian: Bengali, Nepali, Asian Indian, Punjabi, Pakistani, Sri Lankan. Southeast Asian: Burmese, Cambodian, Filipino, Indonesian, Laotian, Hmong, Malaysian, Thai, Vietnamese). Student characteristics include gender, days absent, and an indicator for economic disadvantage. The sample of students in the assessment of teacher overrating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$. Classification of counties into subgroups shares based on ACS self-reported ancestry data from 2007-2013.

To help understand what might be driving results in Table C3, Table C4 provides descriptive information for students in city versus non-city set-

Table C3: Racial Differentials in Teacher Assessments, City vs. Non-city

	Over-rate ($NB > B$) (1)	Under-rate ($NB < B$) (2)
Asian×City	-0.014***	0.009***
	(0.004)	(0.002)
Black×City	-0.009***	0.009***
	(0.001)	(0.001)
Hispanic×City	-0.012***	0.008***
	(0.002)	(0.002)
Native American×City	-0.003	0.012**
	(0.007)	(0.006)
Other×City	-0.009***	0.006***
	(0.003)	(0.002)
Asian	0.047***	-0.030***
	(0.003)	(0.002)
Black	-0.020***	0.016***
	(0.001)	(0.001)
Hispanic	-0.018***	0.018***
	(0.001)	(0.001)
Native American	-0.020***	0.009***
	(0.003)	(0.002)
Other	-0.002	0.003**
	(0.001)	(0.001)
N	11,712,857	13,413,211

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. Omitted teacher race: White teachers. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

tings by race. We see that Asian students in cities have relatively lower socioeconomic and academic outcomes than White peers compared to Asian students in rural, town, or suburban locations. Asians are 16 percentage points more likely to be economically disadvantaged than White peers in cities, while they are only 7 percentage points more likely to be economically disadvantaged in non-city locations. The average lagged math score of Asian students in cities is .12 standard deviations higher than that of White peers, while the corresponding measurement is .26 standard deviations higher in non-city locations. The average lagged reading score of Asian students in cities is .17 standard deviations lower than that of White peers, while the corresponding measurement is .02 standard deviations lower in non-city locations. Overall, information in Table C4 suggests that positive stereotyping towards Asian students may be lower in urban areas because Asians in these areas tend to conform less to the "model minority" stereotype.

Table C4: Descriptive Student Statistics by Race, City vs. Non-city

	White		Asian		Black		Hispanic	
	City	Non-City	City	Non-City	City	Non-City	City	Non-City
Econ disadvantaged	0.21	0.34	0.37	0.41	0.73	0.75	0.83	0.84
	(0.41)	(0.48)	(0.48)	(0.49)	(0.44)	(0.43)	(0.38)	(0.37)
Lagged math score	0.56	0.26	0.68	0.52	-0.46	-0.47	-0.26	-0.21
	(0.93)	(0.91)	(1.11)	(0.99)	(0.90)	(0.88)	(0.93)	(0.89)
Lagged reading score	0.56	0.27	0.39	0.25	-0.41	-0.43	-0.41	-0.35
	(0.90)	(0.91)	(1.09)	(0.98)	(0.92)	(0.90)	(0.97)	(0.93)

Observations are at the student-year level for students in grades 3-8 in math or reading classes between 2007-2013. Lagged test scores are measured as z-scores.

Table C5 examines whether racial gaps in teacher assessment varies across teacher characteristics. Specifically, we assess whether the extent of racial differentials is associated with teacher race, age, and experience. The top panel looks at whether White and Non-White teachers differ in the propensities to over-rate or under-rate Asians, and estimates do not indicate differences across teacher race. Due to the very small number of Asian teachers in our sample, we do not have enough statistical power to look at the effects of having an Asian teacher on assessment outcomes for Asian students. The middle panel assesses whether experience affects teacher's propensities to over-rate or under-rate Asian students relative to White peers and also does not find evidence of differences. Finally, the bottom panel looks at whether teacher age affects assessment outcomes, and we do not find evidence that this characteristic influences assessments of Asian students

Table C5: Racial Differentials in Teacher Assessments by Teacher Characteristics

	Over-rate $(NB > B)$	Under-rate $(NB < B)$
Teacher race		
Asian	0.042***	-0.028***
	(0.002)	(0.001)
Asian× Teacher Non-White	0.005	-0.001
	(0.005)	(0.003)
N	11,797,047	13,495,796
Teacher experience		
Asian	0.042***	-0.024***
	(0.003)	(0.002)
Asian×Teacher experience (10 yr)	0.001	0.001
	(0.002)	(0.001)
N	11,797,047	13,495,796
Teacher age		
Asian	0.041***	-0.025***
	(0.007)	(0.004)
Asian×Teacher age (10 yr)	0.0004	-0.0006
	(0.002)	(0.001)
N	11,731,384	13,430,028

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013. Omitted category: White students. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Teacher experience and teacher age have been rescaled so that a one unit increase represents an increase in experience or age by 10 years, respectively. Student characteristics include controls for gender, economic disadvantage, lagged days absent, and lagged disciplinary infractions. Each panel also includes controls for students of other racial groups (Black, Hispanic, Other race, American Indian) and these controls interacted with the relevant teacher characteristics, although these coefficients are not displayed. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

D Spillover Effects

Table D1: Effect of Exposure to Asian Students, Overall and by Attribute

	Over-rate ($NB > B$)			Under-rate ($NB < B$)		
	All	1 Asian	2+ Asians	All	1 Asian	2+ Asians
Black×Any Asian	-0.005***	-0.005***	-0.007**	0.003**	0.005***	0.001
	(0.002)	(0.002)	(0.003)	(0.001)	(0.002)	(0.002)
Hispanic×Any Asian	-0.005***	-0.006***	-0.004	0.003	0.002	0.006**
	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.003)
Native American×Any Asian	0.004	-0.000	0.019	0.001	0.004	-0.006
	(0.008)	(0.009)	(0.018)	(0.007)	(0.008)	(0.015)
Other×Any Asian	-0.002	-0.004	-0.000	-0.002	-0.000	-0.007
	(0.003)	(0.004)	(0.006)	(0.003)	(0.003)	(0.005)
N	11,785,522	10,614,152	9,706,445	13,496,166	11,789,383	10,945,708
Race×teacher-school-grade-course FE	Y	Y	Y	Y	Y	Y

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. Sample comprises students across all racial groups in grades 3-8 between 2007-2013. The "All" sample includes all classrooms, while "1 Asian" is limited to classrooms that have either zero or one Asian student and "2+ Asians" include classrooms with either zero or at least two Asian students. Any Asian is a binary variable indicating that the classroom had at least one Asian student. The omitted category is White students. Models include interactions between Native American students, and students of other racial or ethnic groups with the Any Asian variable. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, an indicator for economic disadvantage, lagged days absent, and lagged disciplinary infractions. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

Table D2: Effect of Exposure to Asian Students on Contemporaneous Achievement and Behavioral Outcomes

	End-of-Grade z-scores	Days absent	Disciplinary infractions
Black×Any Asian	-0.004	0.024	-0.007
	(0.002)	(0.023)	(0.006)
Hispanic×Any Asian	-0.005	0.043	-0.001
	(0.003)	(0.029)	(0.006)
Native American×Any Asian	-0.018	0.044	-0.007
	(0.011)	(0.131)	(0.006)
Other×Any Asian	-0.000	0.007	-0.001
	(0.005)	(0.048)	(0.006)
Race×Teacher-school-grade-course FE	Y	Y	Y
N	11,027,128	13,359,278	13,359,549

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. Sample comprises students across all racial groups in grades 3-8 in odd-numbered columns and grades 4-8 in even-numbered columns between 2007-2013 in classrooms with zero or one Asian student. Any Asian is a binary variable indicating that the classroom had one Asian student. The omitted category is White students. All specifications include controls for class fixed effects, gender, an indicator for economic disadvantage, lagged days absent, and lagged disciplinary infractions. The first column furthermore includes baseline lagged EOG z-scores.

Table D3: Effect of Exposure to One Black Student

	Over-rate $(NB > B)$ (1)	Under-rate $(NB < B)$ (2)
Hispanic×Any Black	0.004	0.003
	(0.004)	(0.004)
Class FE	Y	Y
Race×teacher-school-grade-course FE	Y	Y
N	2,786,881	3,841,711

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013 and is limited to classrooms that have either zero or one Black student. Any Black is a binary variable indicating that the classroom had one Black student. Omitted category: White students. Models include interactions between Asian students, Native American students, and students of other racial or ethnic groups with the Any Black variable. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

Table D4: Effect of Exposure to Black Students, by Achievement and SES

	Over-rate $(NB > B)$	Under-rate $(NB < B)$
Panel A: By Achievement		
Hispanic×Black Lagged Z-score - Quartile 1	0.004	0.001
	(0.005)	(0.005)
Hispanic×Black Lagged Z-score - Quartile 2	0.000	-0.005
	(0.007)	(0.006)
Hispanic×Black Lagged Z-score - Quartile 3	0.000	0.002
	(0.008)	(0.007)
Hispanic×Black Lagged Z-score - Quartile 4	0.001	0.007
	(0.012)	(0.008)
N	2,233,994	3,130,482
Panel B: By Socioeconomic Status		
Hispanic×High SES Black	0.009	0.004
	(0.006)	(0.005)
Hispanic×Low SES Black	0.001	0.001
	(0.005)	(0.004)
N	2,785,263	3,839,471
Race×Teacher-school-grade-course FE	Y	Y

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. Sample comprises students across all racial groups in grades 3-8 between 2007-2013 in classrooms with zero or one Black student. Any Black is a binary variable indicating that the classroom had one Black student. Lagged Z-score is the Black student's standardized lagged z-score normalized within the population of all students in a given grade and year. Quartiles are defined using the statewide sample of students at the year, grade, and subject level (e.g. quartile 1 is the 25th percentile or below). Low (high) SES assumes a value of 1 if the Black student is (not) economically disadvantaged, defined as being eligible for free or reduced price lunch. The omitted category is White students. Models include Native American and students of other racial or ethnic groups and their interactions with the Any Black, Black Lagged Z-score quartiles, and Low SES Black variables. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage.

Table D5: Effect of Exposure to One Hispanic Student

	Over-rate $(NB > B)$ (1)	Under-rate $(NB < B)$ (2)
Black×Any Hispanic	0.001	-0.002
	(0.002)	(0.002)
Class FE	Y	Y
Race×teacher-school-grade-course FE	Y	Y
N	4,850,776	6,098,442

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. Sample comprises students in grades 3-8 between 2007-2013 and is limited to classrooms that have either zero or one Hispanic student. Any Hispanic is a binary variable indicating that the classroom had one Hispanic student. Omitted category: White students. Models include interactions between Asian students, Native American students, and students of other racial or ethnic groups with the Any Black variable. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage. The sample of students in the assessment of teacher over-rating includes those with $B \in \{1, 2, 3\}$. The sample of students in the assessment of teacher under-rating includes those with $B \in \{2, 3, 4\}$.

Table D6: Effect of Exposure to Hispanic Students, by Achievement and SES

	Over-rate ($NB > B$)	Under-rate ($NB < B$)
Panel A: By Achievement		
Black×Hispanic Lagged Z-score - Quartile 1	0.009***	-0.004
	(0.003)	(0.003)
Black×Hispanic Lagged Z-score - Quartile 2	0.001	0.002
	(0.004)	(0.003)
Black×Hispanic Lagged Z-score - Quartile 3	0.002	-0.002
	(0.004)	(0.003)
Black×Hispanic Lagged Z-score - Quartile 4	-0.010*	0.003
	(0.006)	(0.004)
N	3,943,586	5,008,713
Panel B: By Socioeconomic Status		
Black×High SES Hispanic	-0.004	0.002
	(0.003)	(0.003)
Black×Low SES Hispanic	0.003	-0.004*
	(0.002)	(0.002)
N	4,848,376	6,094,714
Race×Teacher-school-grade-course FE	Y	Y

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. Sample comprises students across all racial groups in grades 3-8 between 2007-2013 in classrooms with zero or one Hispanic student. Any Hispanic is a binary variable indicating that the classroom had one Hispanic student. Lagged Z-score is the Hispanic student's standardized lagged z-score normalized within the population of all students in a given grade and year. Quartiles are defined using the statewide sample of students at the year, grade, and subject level (e.g. quartile 1 is the 25th percentile or below). Low (high) SES assumes a value of 1 if the Hispanic student is (not) economically disadvantaged, defined as being eligible for free or reduced price lunch. The omitted category is White students. Models include Native American and students of other racial or ethnic groups and their interactions with the Any Hispanic, Hispanic Lagged Z-score quartiles, and Low SES Hispanic variables. All specifications include controls for observable student characteristics, class fixed effects, and raw end-of-grade test score fixed effects interacted with subject, grade, and year. Student characteristics include gender, lagged days absent, lagged disciplinary infractions, and an indicator for economic disadvantage.