Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1361

# Effects of a non-traditional teacher preparation program on non-test outcomes: evidence from relay graduate school of education in New York City

Soobin Kim
University of Wisconsin – Madison

This study examines the effects of a non-traditional teacher preparation program, the Relay Graduate School of Education, on non-test outcomes for New York City public school students in Grades 3–8. By controlling for student and school fixed effects, I use plausibly random variation in Relay teacher assignments within students over time to identify causal Relay program effects. Results indicate that Relay-trained teachers are more effective at improving student attendance and marginally effective at reducing suspensions compared to non-Relay teachers. The program shows a particular impact on students of color, male students, and those from economically disadvantaged backgrounds.

VERSION: December 2025

Suggested citation: Kim, Soobin. (2025). Effects of a non-traditional teacher preparation program on non-test outcomes: evidence from relay graduate school of education in New York City. (EdWorkingPaper: 25-1361). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/j340-jq26

# Effects of a non-traditional teacher preparation program on non-test outcomes: evidence from Relay Graduate School of Education in New York City

Soobin Kim[1]

## Abstract

This study examines the effects of a non-traditional teacher preparation program, the Relay Graduate School of Education, on non-test outcomes for New York City public school students in Grades 3–8. By controlling for student and school fixed effects, I use plausibly random variation in Relay teacher assignments within students over time to identify causal Relay program effects. Results indicate that Relay-trained teachers are more effective at improving student attendance and marginally effective at reducing suspensions compared to non-Relay teachers. The program shows a particular impact on students of color, male students, and those from economically disadvantaged backgrounds.

Keywords: Student outcomes, attendance and suspension, teacher preparation

## 1. Introduction

Over the past two decades, research on alternative teacher preparation programs has grown substantially, with particular emphasis on Teach For America (TFA). A robust literature has evaluated TFA's impact on standardized test scores using both experimental studies (Antecol, Eren, & Ozbeklik, 2013; Clark & Isenberg, 2020; Glazerman, Mayer, & Decker, 2006; Lovison, 2022; Penner, 2016; 2021) and quasi-experimental analyses of administrative data (An & Koedel, 2021; Backes & Hansen, 2018; Cochra-Smith & Reagan, 2021; Henry et al., 2014; Xu, Hannaway, & Taylor, 2011). However, this research has predominantly focused on academic achievement measured by standardized tests, leaving a significant gap in our understanding of how alternative certification pathways influence other important dimensions of student success.

Understanding the effects of specific teacher preparation programs is particularly important given that broader research often finds general teacher characteristics, including certification status and advanced degrees, are not consistently predictive of student outcomes (Clotfelter et al., 2007; Goldhaber & Brewer, 2000; Kane et al., 2008). This suggests that how teachers are prepared, rather than the credentials they hold, may be the key driver of effectiveness. Student non-test outcomes have been shown to be a critical factor driving academic achievement and other long-term outcomes (Blazar & Kraft, 2017; Heckman et al., 2006; Jackson, 2018). For example, higher attendance is associated with increases in student achievement, high school graduation, and college enrollment (Aucejo & Romano, 2016; Gershenson, 2016; Gershenson et al., 2017; Gottfried, 2009; 2011; Jackson, 2018; Liu et al., 2021; Tran & Gershenson, 2021). Similarly, suspensions contribute to achievement gaps (Morris & Perry, 2016; Pearman et al., 2019) and to the likelihood of involvement with the criminal justice system (Bacher-Hicks et al., 2024; Curtis, 2013; Davison et al., 2022). Despite the

[1] Soobin Kim, Center for Demography of Health and Aging, University of Wisconsin – Madison, skim83@wisc.edu. 4321 William H. Sewell Social Sciences, 1180 Observatory Dr. Madison, WI 53706.

importance of these non-test outcomes, little research has examined how alternative teacher preparation programs affect them, particularly for programs beyond TFA.

This study examines the effects of teachers trained at the Relay Graduate School of Education: a high-profile, non-profit, accredited teacher preparation program with locations across the United States. I investigate whether having a Relay-trained teacher improves student non-test outcomes and whether the effects of Relay teachers differ by student characteristics. Unlike TFA – which primarily recruits recent college graduates for short-term teaching commitments with minimal in-service training – Relay employs a fundamentally different approach to teacher preparation. Founded in 2011 by leaders from high-performing charter networks, Relay specifically targets teachers already working in schools, providing sustained, practice-based training integrated with classroom experiences (Mungal, 2016). The program emphasizes `deliberate practice` of specific teaching techniques, culturally responsive pedagogy embedded throughout the curriculum rather than as standalone modules, and comprehensive classroom management strategies. Relay's model centers on developing teachers' abilities to create productive learning environments and build strong relationships with students from diverse backgrounds – skills that may be particularly relevant for improving student engagement and reducing disciplinary incidents. This distinctive approach to teacher preparation, with its emphasis on practical application and cultural responsiveness, offers a compelling context for examining effects on non-test outcomes like attendance and suspensions.

To answer these questions, I use longitudinal administrative data from the New York City Department of Education (NYCDOE), tracking students in Grades 3–8 between the 2013–2014 and 2018–2019 school years. The rich student-level data linked to teacher data are well suited for this analysis because they allow me to observe both changes in student non-test outcomes and exposure to Relay teachers. I identify the effects of Relay teachers that are plausibly unbiased conditional on student fixed effects.

I present three key findings. First, I find that Relay teachers instruct more students of color, students with disabilities, students who are English language learners (ELLs), students from economically disadvantaged (ED) backgrounds, and students in higher grades (i.e., Grades 6–8); and these students have more days of prior absences and suspensions compared to non-ELL, non-ED, and lower grade students.

Second, I find that Relay teachers are marginally more effective than non-Relay teachers in improving student non-test outcomes. I compare students' attendance and suspension outcomes in years when they took more courses taught by Relay teachers with their own non-test outcomes in years when they took fewer courses with Relay teachers. Relay teachers are more effective in reducing absences, but the magnitude is small. Relay effects on suspension outcomes are marginally significant in principal suspensions. For example, when students were taught entirely by Relay teachers, they were likely to have 0.4 fewer absences (0.048 standard deviations) in a school year and a 20% reduction in the number of principal suspensions. I do not find any detectable effect of Relay teachers on relatively extreme and infrequent behaviors, such as chronic absenteeism and superintendent suspensions.

Third, I find that Relay teachers' effects on student attendance and suspension outcomes are more pronounced among students of color and male students. Relay teachers are also effective in reducing absences among ED students, though their effects on suspensions are less clear. In sum, Relay teachers are more likely to teach students of color and historically marginalized students in need of the most support, and those are the students for whom Relay teachers have the largest effects.

This paper makes two major contributions to the literature on teacher preparation. First, it extends the limited empirical evidence on how alternatively prepared teachers affect non-test outcomes that strongly predict long-term student success. While standardized test scores remain important metrics, attendance and disciplinary incidents also significantly predict high school completion, college enrollment, and future earnings (Gershenson, 2016; Jackson, 2018). Recent research has started to examine these broader outcomes, with Backes and Hansen (2018; 2024) finding modest improvements in student attendance and reductions in suspensions attributable to TFA teachers. By explicitly focusing on attendance and suspensions, this study provides a more comprehensive understanding of teacher effectiveness, particularly given the disproportionate impact of absenteeism and disciplinary incidents on historically marginalized student populations.

Second, this study provides the first rigorous evaluation of the Relay Graduate School of Education, an expanding alternative certification program distinct from TFA in structure and pedagogical approach. Despite its growth and presence across multiple states, Relay has received limited empirical scrutiny. Unlike TFA – which primarily recruits recent college graduates for short-term teaching commitments with minimal in-service training – Relay targets teachers already employed in schools, providing sustained, practice-based training integrated with classroom experiences (Mungal, 2016). Relay's model emphasizes practical instructional strategies, culturally responsive teaching embedded throughout its curriculum, and strong classroom management – features well-positioned to enhance student engagement and behavior. As alternative certification pathways diversify beyond TFA, rigorous evaluations of different program models become essential for informing effective policy and practice.

The findings of this study demonstrate that Relay teachers significantly improve student attendance and reduce suspensions compared to non-Relay teachers, with particularly pronounced effects among students of color and male students. These heterogeneous effects are especially meaningful given persistent racial and gender disparities in school disciplinary practices and attendance. By highlighting Relay teachers' effectiveness specifically with student populations most adversely impacted by these disparities, this study provides critical insights for policies, aiming to address educational inequities through teacher training.

As policymakers increasingly seek to improve student outcomes beyond standardized testing, understanding the impacts of diverse teacher preparation programs becomes crucial. This research not only enriches our knowledge of Relay's effectiveness but also broadens the discourse on teacher preparation by emphasizing outcomes that predict long-term student success.

## 2. Program description: The Relay Graduate School of Education

The Relay Graduate School of Education was founded in 2011 by educational leaders from high-performing charter networks, including KIPP, Achievement First, and Uncommon Schools. Relay spun off from Hunter College CUNY to address critical gaps in traditional teacher preparation, particularly for high-needs schools in underserved communities. The founders identified a need for more practice-focused, classroom-ready teachers capable of delivering high-quality instruction and culturally responsive practices to diverse student populations in challenging educational environments. Relay has grown to operate teacher preparation programs in 10 states plus the District of Columbia, offering one and two-year advanced certification, master's degrees, residency programs, and leadership development.

Unlike traditional university-based teacher education programs – which typically require extensive coursework and pedagogical theory prior to classroom teaching experience – Relay integrates pedagogical theory, content, and clinical practice simultaneously. Traditional programs usually provide student-teaching placements only after completing theoretical courses. In contrast, many Relay candidates are already employed as teachers under temporary or provisional certification, allowing immediate application and iterative feedback on pedagogical techniques within their own classrooms.

Additionally, whereas traditional programs often isolate critical teaching competencies such as classroom culture and culturally responsive pedagogy into separate, specialized coursework, Relay treats these elements as integral and weaves them comprehensively throughout its entire curriculum. By embedding cultural responsiveness and classroom management deeply into every aspect of its training, Relay positions these competencies as foundational rather than supplemental skills. Relay's continuous feedback mechanisms – including video analysis, coaching, and peer feedback – create a structured practice environment rarely found in conventional teacher preparation programs.

These distinctive features provide theoretical reasons to expect different outcomes from Relay teachers. The program's emphasis on embedded classroom management techniques and culturally responsive teaching practices may be particularly effective for improving student engagement and reducing disciplinary incidents. Because Relay prepares teachers specifically for the contexts in which they already serve, this targeted preparation potentially leads to better outcomes for historically marginalized student populations.

Studying Relay's effectiveness is particularly important given its rapid expansion to multiple states and its position as a prominent alternative to both traditional certification and its partnership with Teach For America. Understanding whether and how Relay's model affects student outcomes provides valuable insights for policy and practice in teacher preparation.

## 2.1 Recruitment

Relay employs several strategies to recruit student teacher candidates. Its primary approach involves developing strong partnerships with local public and charter schools that seek diverse teacher candidates and prioritize higher teacher retention, performance, satisfaction, and student outcomes in high-needs schools. For example, Relay formerly partnered with the NYC Teaching Fellows program aiming to recruit diverse teaching candidates who then enrolled in Relay's master's program. Furthermore, partner schools located in underserved communities and/or communities of color may recommend teachers to Relay for teacher certification programs. Another example of a targeted partner is Teach for America. While these programs focus on recruiting teacher candidates, Relay and other educator preparation programs provide them with clear pathways to master's degrees and certification.

Relay offers a high degree of support for students in their program, based on a culturally responsive student advising model where students are offered extensive support if they are struggling (Bowes, 2017). In the past, the program's affordability was another characteristic that supported teachers of color who might have been unable to attend traditional teacher preparation programs due to financial constraints. Relay endeavors to make the program accessible by assisting students in obtaining financial support and providing discounted tuition for certain partners, such as NYC Teaching Fellows.

## 2.2 Training

Hallmarks of Relay's educator preparation program include a practice-based approach centered on rounds of observation and feedback; faculty who are highly effective former teachers and leaders; and a content rich curriculum that weaves cultural responsiveness throughout all courses rather than as a standalone unit. Student advising is culturally responsive and inclusive, as well, to further model how Relay's students can provide similar support to their Birth–12 students.

Each of the above hallmarks is meant to enhance the effectiveness of Relay-trained teachers. As part of their practice-based approach, Relay students practice key teaching components, such as developing a strong teacher presence and a safe and productive classroom environment; effective lesson planning; and providing individual interventions to better meet student needs. Through rigorous practice, they enhance their capacity to build relationships with their Birth–12 students and caregivers. Similarly, faculty with real-world experiences are able to provide examples of and feedback on potential solutions to challenges in the classroom.

Incorporating cultural responsiveness throughout the program helps teachers see how it may be applied in different parts of their work. Another curricular element is providing clear rationales for best practices so that Relay students can more clearly understand application and context. Relay believes that providing the `why` with the `what` equips their teachers with knowledge on how to approach a given situation.

Finally, by providing culturally responsive and inclusive advising, Relay hopes to serve the whole student, giving them human-centered, individualized support throughout the program (Bowes, 2017). This type of support has been shown to bolster the academic achievement of teacher candidates who are Black, Indigenous, or other people of color (Blazar, 2021). Relay's NYC campus also takes a collaborative, case management approach to advising that utilizes as many available Relay resources as possible to support students. As stated above, a key benefit of culturally responsive and inclusive advising is modeling how Relay's students can build relationships with their students and meet their needs.

In NYC, Relay recruited and trained new teachers to prepare them to teach in high-needs schools, including charter schools, within the NYCDOE. In the broader context of teacher preparation literature, it is essential to recognize the dual pathways – recruitment and training – through which Relay-trained teachers can influence student outcomes. This distinction is highlighted in studies by Goldhaber et al. (2013), Koedel et al. (2015), and Boyd et al. (2009). The Relay program effects may stem from the specialized training they receive or the rigorous selection process employed by Relay; however, the limitations of the current research design and available data prevent the differentiation of these effects. A more comprehensive analysis would require observation of a significant number of teachers before they enroll in Relay. Such an approach would offer insights into the extent to which Relay training contributes to enhancing teacher quality – an avenue for future exploration.

## 3. Data

To evaluate the effects of the Relay program on student attendance and suspension outcomes, I use multiple data sets covering the 2013–2014 through 2018–2019 school years. First, I obtained a list of teachers who graduated from the Relay Graduate School of Education. Second, I received course-taking data from the NYCDOE, which include student ID, the number of credits attempted/earned, the course section, teacher ID, school ID, and course start and end dates. I used a data crosswalk available on the New York State Education

Department's website,[2] which lists all state course codes and their subjects, to identify the subject of each course in the NYCDOE data. The official state course ID variable was used to identify the courses to which the students were assigned in each school year. I used these data to match students to teachers in specific subjects.

Lastly, I obtained student-level data from the NYCDOE on student demographics, yearly and total attendance and suspension, and state assessment scores. The demographic records contain information on student race/ethnicity, gender, ELL status, ED status, and receipt of IEP services.[3] Course-level data on absences and suspensions were not available.

### 3.1 Measures

I focus on several key measures in this analysis, including student non-test outcomes, exposure to Relay-trained teachers, and student demographic characteristics. I also include time-varying controls at the school, teacher, and classroom levels to account for potential confounding factors.

First, for attendance outcomes, the average number and standard deviation of absences differ by grade and year. Consequently, interpreting raw absence counts consistently across different grades and years can be challenging, as the meaning or impact of a single day of absence may vary depending on the student's grade level. To address this and to facilitate the comparison of effect magnitudes, I normalize the number of absences by grade and year, transforming them to have a mean of zero and a standard deviation equal to one within each grade-year cohort. This normalization allows the estimated coefficients from the models to be interpreted as effect sizes, representing fractions of a standard deviation change in absences. Another attendance indicator used is whether a student is chronically absent, defined as missing 10% or more of school days.

Second, for suspension outcomes, suspensions are recorded for each suspended student at a school in a given year. There are two different types of suspension outcomes in the data: principal suspensions (one to five school days) and superintendent suspensions (more than five days). Superintendent suspensions typically involve more serious infractions and occur less frequently. Due to data limitations, I cannot distinguish between superintendent suspensions of different durations or reasons. Given their rarity and the fact that they represent extreme behaviors, the primary analyses focus on principal suspensions, though I also report results for total suspensions (both principal and superintendent combined). Similar to absences, the frequency and distribution of suspensions can vary considerably by grade and year. To account for these variations and to allow for the interpretation of effects in terms of standardized units, the number of suspensions (both principal and total) is normalized by grade and year, transforming them to have a mean of zero and a standard deviation of one within each grade-year cohort. This approach enables me to report effect sizes and compare the relative impact of Relay teachers on suspensions despite differences in baseline rates and variability across grades and years.

Third, I construct a Relay dosage variable to measure student exposure to Relay-trained teachers. For each course, the Relay indicator variable is assigned a `1` if a student received course instruction from a Relay teacher at a specified school during the school year. Although

[2] http://www.p12.nysed.gov/irs/courseCatalog/home.html

[3] The NYCDOE demographic data dictionary states that, to comply with legal requirements, the DOE can no longer release the meal code for individual students. Instead, an indicator is provided that signifies whether a student qualifies for a free or reduced lunch, or attends a universal feeding school.

students in the sample took as many as 13 courses taught by Relay teachers, fewer than 4% of students across all years took at least 1 course taught by Relay teachers.

Given that Relay's effects on outcomes could be larger among students with greater exposure to Relay teachers, I create a *Relay dosage variable* for each student in a given year, directly proportional to the number of courses in which a specific teacher instructed a specific student. This dosage variable quantifies students' exposure to Relay teachers within a given academic year, calculated as the proportion of their classes taught by Relay-trained teachers. If a student took more courses with one teacher, that teacher would likely spend more time with that student and exert more influence on their outcomes. For example, suppose that a student took four courses taught by two teachers, X and Y. Non-Relay Teacher X taught the student in three courses (identifiable by unique course IDs), and Relay Teacher Y taught the student in one course (distinct from Teacher X's courses). Non-Relay Teacher X would be assigned a weight of ¾ and Relay Teacher Y would be assigned a weight of ¼ for this specific student. As Relay Teacher Y taught the student 25% of the instructional time, the Relay dosage of Teacher Y would be 0.25. This continuous measure allows me to precisely capture incremental impacts of Relay exposure on student outcomes. In cases of co-teaching, the dosage variable is adjusted accordingly. For example, if a course was co-taught by one Relay teacher and one non-Relay teacher, the instructional time for that course would be split, with only half contributing to the Relay dosage.

Fourth, I include student demographic variables in the analysis: student race/ethnicity, gender, ELL status, ED status, and receipt of an IEP. Some of these student program variables are *time variant*, meaning that students can be identified as ED, an ELL, or as receiving an IEP in some years and not in others. For example, a student may be identified as an ELL in elementary school but not in middle school. In addition to these program variables, I calculate *time-invariant* program variables to reflect whether students were (a) ever, (b) always, or (c) never identified as ED, an ELL, or as qualifying for an IEP across years.

Finally, the analysis incorporates three sets of time-varying controls – school-level, teacher-level, and classroom-level – to account for potential confounding factors. These controls are crucial for addressing potential non-random sorting of students to Relay teachers and for isolating the specific effects of the Relay program from other factors correlated with both teacher assignment and student outcomes. School-level controls are included to account for time-varying school characteristics that may be associated with student outcomes and teacher assignment. These controls include the school-wide average of lagged student test scores (standardized math and ELA test scores) and the school-wide proportions of students with various demographic characteristics (Black, Hispanic, Asian, White, and other races; ELL, IEP, and Economically Disadvantaged status; and male gender).

Teacher-level controls include demographic characteristics and professional experience. Specifically, I control for proportions of a student's instructional time taught by teachers of a particular race/ethnicity (Asian, Black, Hispanic, Multi-racial, Other/Unknown, White), the proportion of a student's instructional time taught by female teachers, and the average years of experience of a student's teachers, weighted by instructional time. These controls help to account for variations in student outcomes that may be attributable to teacher characteristics other than their Relay training. Classroom-level controls capture the characteristics of a student's peers within their classrooms that may be correlated with outcomes. These controls include the average of lagged peer test scores (standardized math and ELA test scores) within a student's classrooms, weighted by instructional time, and the proportions of peers with various demographic characteristics within a student's classrooms (Economically Disadvantaged, ELL, IEP status; female gender; and Asian, Black, Hispanic, and White race), weighted by instructional time. These peer characteristics are calculated as averages across all of a student's courses in a given year, weighted by instructional time.

### 3.2 Sample

I use data on all students in Grades 3–8 attending public school in NYC between 2014 and 2019 and make a series of sample restrictions.[4] First, I focus on four academic subjects: math, English language arts (ELA), social studies, and science. This restriction excludes 51,470 courses, leaving 59,801 for inclusion. By restricting the data to these key academic courses, I implicitly assume that each teacher in these subjects (including co-teaching teachers) contributes equally to the changes in students' outcomes, though I acknowledge that teachers in non-core subjects could also affect student outcomes. Additional analyses indicate that results are robust to the inclusion of additional subjects. Second, I drop students whose total number of school days (sum of the number of days absent and number of days present) was below 175 days for a school year or the prior school year. I further drop students whose teacher–student linkage or demographic information is not available (1.96% of the sample). For students whose demographic identifications changed across years (approximately 0.36% for gender and 2.53% for race/ethnicity), I assigned time-invariant classifications based on the most frequently recorded identity, prioritizing the most recent identification in cases of equal frequency. The analytic sample contains 2.2 million student-year observations on about 898,000 unique students attending over 1,200 schools from 33 NYC geographic districts. These data cover roughly 800,000 unique classrooms taught by more than 54,000 unique teachers.

The summary statistics for teachers and students are presented in Table 1. Panel A shows the demographic characteristics of Relay-trained and non-Relay-trained teachers. Relay-trained teachers account for 0.7% of all teachers in NYC and 4% of first-year teachers. Relay-trained teachers are less likely to be female but more likely to be teachers of color. As the Relay program started in 2012, Relay-trained teachers have fewer years of experience compared to non-Relay-trained teachers. The number of Relay-trained teachers has increased over time, from 118 in 2014 to 342 in 2018, comprising approximately 4% of all first-year teachers and 1% of all NYC public school teachers (Appendix Table 1). Appendix Table 2 shows that among first-year teachers, Relay teachers are more diverse, with 56% being non-White compared to 42% among non-Relay teachers, and include a higher proportion of male teachers (30% versus 21%)

Panel B presents students' demographic characteristics. In addition to the overall sample statistics (Column 1), I present the statistics of students who took at least one course taught by Relay teachers (Column 2) and of students who took no courses with Relay teachers (Column 3) in a given year. In Column 4, I also present the difference in means between students taught by Relay teachers (Column 2) and those never taught by Relay teachers (Column 3). Seventeen percent of the students in this sample are White, 23% are African American, 41% are Hispanic,

[4] This study focuses on students in Grades 3 through 8 for methodological reasons. High school settings (Grades 9-12) present distinct analytical challenges due to complex scheduling and data limitations that complicate the attribution of attendance and behavioral outcomes to specific teachers using my dosage measure. Exploratory analyses (not reported) revealed substantially different and inconsistent patterns in high schools compared to grades 3-8. These results are difficult to interpret definitively but are likely due to structural differences in high school attendance patterns (e.g., class-skipping) which affect the reliability of my data and methods in that context. Focusing on grades 3–8 ensures more precise estimation where my identification strategy is more robust. Future research with more detailed high school data is needed.

18% are Asian, and 1% are Other. Forty-nine percent of the students are female. Seventy-three percent of the students are ED, and 12% have limited English proficiency. Twenty-two percent have an IEP.

In terms of students' demographic characteristics, Relay teachers instruct more students of color (i.e., higher percentages of Black and Hispanic students), more ELL students, more ED students, and more students receiving an IEP. Thus, if there are differential rates of student absences and suspensions across these characteristics (for example, more absences for students of color), unconditional comparisons of students' absences and suspensions between those two groups is likely to underestimate Relay program effects.

These descriptive differences suggest non-random sorting of Relay teachers, but simple mean differences don't fully capture the extent of this selection. To quantify this pattern more precisely, I constructed course-level transcript data for Grades 3–8 between 2014 and 2019, restricting the sample to four key academic subjects taught by first-year teachers. Regressing a Relay teacher assignment indicator on student characteristics with grade, year, and district fixed effects reveals that male students, Black students, Hispanic students, Asian students, and economically disadvantaged students are significantly more likely to be assigned to Relay teachers (Appendix Table 3). This clear evidence of non-random sorting underscores the importance of using student fixed effects in the main analysis to address potential selection bias.

Panels C and D show averages of key non-test outcomes as well as those from the prior year. On average, students are absent about 10 days per year, and the standard deviation of about 9 indicates that there is substantial variation across student years in the sample. Comparing Columns 2 and 3 shows that there are statistically significant differences between students who are taught by Relay teachers and those who are not, in terms of their outcomes. For example, students of Relay teachers miss an additional 1.9 days, compared to students of non-Relay teachers. In addition, students taught by Relay teachers receive twice as many principal and superintendent suspensions, compared to students taught by non-Relay teachers. If I compare their prior year outcomes, students taught by Relay teachers had lower test scores and more absences in the prior year than students never taught by Relay teachers. Thus, Relay teachers appear to be assigned, on average, to lower-achieving students and to students who are more likely to be absent and get suspended.

### 3.3 Differences in non-test outcomes by demographic characteristics

I examine variations in student non-test outcomes by grade, gender, and race/ethnicity, as well as ELL, ED, and IEP status. Appendix Table 4 indicates that absences vary among grades and genders, with racial differences being more pronounced – Black and Hispanic students had about two more absences than White students, while Asian students had fewer. Students with an ED status and those with an IEP had notably more absences.

Appendix Table 5 examines variations in principal suspensions across student demographic characteristics. Superintendent suspensions are excluded due to their rarity. Boys have double the suspensions compared to girls, and racial disparities are evident: Black students have the highest suspension rates, followed by Hispanic, White, and Asian students. Always-ED students are suspended three times more than Never-ED students. IEP students have twice the suspension likelihood compared to non-IEP students, and minimal differences are observed for ELL students. Elementary students have significantly fewer suspensions than middle schoolers.

Overall, significant disparities in non-test outcomes exist across race, ED, and IEP status. Given that Relay teachers often teach students of color and historically marginalized groups, these disparities must be accounted for, to accurately estimate Relay's effects.

## 4. Method

I use longitudinal student-level data linked to teachers to estimate the effectiveness of Relay program teachers in improving non-test outcomes. As previously described, Relay teachers are different from non-Relay teachers with respect to the students they teach. Relay teachers are more likely to teach students of color and those with historically marginalized backgrounds, who are more likely to be absent or suspended, with lower prior test scores (see Appendix Tables 4 and 5). This indicates that teacher assignment is not random, which, if ignored, is likely to bias the estimate. For example, if Relay teachers are systematically assigned to students with a greater tendency to be absent or suspended, estimated Relay effects are likely to be biased downward (i.e., underestimated). In addition, if Relay teachers are assigned to courses with students who have differences in unobservable factors (e.g., student motivation or academic ability), then cross-sectional approaches that do not account for the non-random matching of teachers and students are likely to estimate biased Relay program effects.

To address the non-random sorting of teachers and students, both across and within schools, I use student fixed-effects models, which incorporate both changes in non-test outcomes and exposure to Relay teachers over time. I additionally control for school fixed effects, grade fixed effects, and any time-varying student characteristics, including ED, ELL, and IEP status. Relay teacher program effects on student non-test outcomes are identified by estimating the following model:

$$A_{ist} = \alpha + \beta Relay_{it} + \gamma X_{it} + \mu_i + \omega_s + \pi_g + \epsilon_{it} \qquad (1)$$

where $A_{ist}$ is the non-test outcome for student *i* attending school *s* in year *t*; $Relay_{it}$ is the fraction of courses that student *i* was taught by Relay teachers in year *t*; $X_{it}$ is a vector of time-varying student characteristics (including ED, ELL, and IEP status); $\mu_i$ are student fixed effects; $\omega_s$ are school fixed effects; $\pi_g$ are grade fixed effects; and $\epsilon_{it}$ is the random error term.

The parameter of interest, $\beta$, is the effect of being taught by Relay-trained teachers on student non-test outcomes. Note that this parameter estimates the average effect of the Relay program, rather than the contribution of individual Relay teachers. The empirical strategy utilizes within-student variation in Relay exposure, comparing individual student outcomes during academic years with higher Relay teacher dosage against the same student's outcomes during years with lower or no Relay exposure. This approach identifies Relay effects by examining how a student's outcomes change when their exposure to Relay-trained teachers varies over time. By incorporating student fixed effects, I control for time-invariant student characteristics that might otherwise bias the estimates, allowing for a more precise isolation of the program's impact. School fixed effects control for non-random sorting of teachers across schools, and grade fixed effects account for time-varying differences in outcomes across grades. Thus, I examine the relationship between (a) within-student variation in exposure to Relay teachers and (b) student non-test outcomes, controlling for differences in outcomes across schools and grades. With these controls, the model identifies Relay effects, provided that teacher assignments are not associated with other time-varying, unobserved determinants of non-test outcomes. I acknowledge that this is a strong assumption and am not claiming that this study achieves perfectly unbiased causal interpretation in the presence of such unobservables (see Rothstein, 2010). Therefore, the estimated effects should be interpreted as plausibly causal, conditional on the included fixed effects and controls.

A critical assumption underlying the identification strategy is that, conditional on student fixed effects, the assignment of students to Relay teachers is plausibly random. To evaluate this assumption, I examine how the likelihood of a student being assigned to a Relay teacher changes with the inclusion of progressively detailed fixed effects. Specifically, I estimate regression models using the Relay dosage as the dependent variable and student, classroom, and school characteristics as predictors (Table 2).

Column 1 in Table 2 shows that without additional controls, there is evidence of non-random sorting, with Black and Hispanic students each approximately 0.8 percentage points more likely to have Relay teachers than their White peers, and economically disadvantaged students 0.3 percentage points more likely. When school and classroom controls are added in Columns 2 and 3, these differences become much smaller and mostly insignificant. The addition of school fixed effects (Column 4) further reduces the magnitudes of these coefficients, though students with IEP and ELL designations still show small but statistically significant differences in Relay teacher assignment.

Most importantly, Column 5 demonstrates that when student fixed effects are introduced, all coefficients related to student demographics – including time-varying characteristics like ELL and IEP status – become statistically insignificant and nearly zero in magnitude. Among 24 classroom and school-level predictors (not all shown), only one retains marginal significance at the 5% level after including student fixed effects.

These findings strongly support the validity of using student fixed effects in the main analysis, as they effectively address potential non-random sorting concerns. The evidence suggests that, conditional on student fixed effects, variation in exposure to Relay teachers can be considered plausibly random, strengthening the causal interpretation of the estimated effects.

## 5. Effects on Non-test Outcomes

Table 3 summarizes the coefficients estimated in Equation (1), by outcome (in panels) and by specification (in columns). Each outcome is estimated separately by specification, which includes progressively richer sets of controls, up to and including school fixed effects. Only the estimated coefficient on Relay is reported. Each column includes controls for student demographic characteristics and grade fixed effects. Standard errors are clustered at the school level, as teachers and students are nested in schools. Clustering at the school level accounts for correlations between students taught by the same teacher and school. Recent methodological research by von Hippel et al. (2016) indicates that standard error estimates for teacher preparation programs can be biased and volatile, but that clustering at higher levels (teacher, school, or district) exhibits little bias for large teacher preparation programs. Following guidance from Cameron and Miller (2015) and Abadie et al. (2023), I cluster at the school level, as robust standard errors may be too small, while more conservative clustering approaches may be unnecessarily restrictive.

### 5.1 Relay effects on attendance outcome

Column 1 in Panel A indicates that students who took all of their courses with Relay teachers had additional absences by 0.084 standard deviations, compared to those who took no courses with Relay teachers. Column 2 shows results from a model that controls for time-varying school-level and classroom-level averages of student race and ethnicity, gender, ED, ELL, IEP, and lagged performance (e.g., number of days absent, standardized math test scores, and standardized ELA test scores); here, the estimated coefficient on the Relay indicator is –0.015, which is not statistically significant.

In Column 3, I introduce several teacher-level average characteristics, such as race, gender, and number of years of experience. The estimated coefficient on the Relay variable changes from -0.015 to -0.047 and is marginally statistically significant. Next, I include school fixed effects in the model. The current data set includes both elementary and middle school students, many of whom have been observed in both settings. Additionally, Relay teachers predominantly serve schools with a higher population of students of color. This trend raises the concern that if these schools implement more exclusionary discipline practices, it might potentially bias the estimates. Results in Column 4 indicate that the inclusion of school fixed effects reduces the estimated coefficients from -0.047 to -0.038.

Finally, when I additionally control for student fixed effects, as shown in Column 5, the estimated coefficient changes to -0.040, which is significant at the 5% level. Considering that one standard deviation in the number of absences is 9.7, this suggests that students who were taught exclusively by Relay teachers (100% of their courses) have 0.4 fewer absences compared to when they have no courses with Relay teachers (0% of their courses).

Although the estimated coefficients illustrate the effects of transitioning from 0% to 100% Relay teachers, it is pertinent to note that Relay teachers only constituted approximately 1.5% of NYC's teacher workforce in 2019. As such, recruiting Relay teachers exclusively across all schools is not feasible. We can contextualize the estimated coefficients by providing a scaling factor rather than focusing on the effect of having all Relay versus non-Relay teachers. If students took 15% of their courses from Relay teachers, compared to the sample average of 1.5% (an increase of 13.5 percentage points, representing roughly 1 standard deviation in the rate of exposure to Relay teachers), the coefficient from Column 5 of Table 3 suggests that the share of Relay teachers increased from 1.5% to 15%, resulting in a reduction in absences by 0.054 (–0.40 × [0.15 – 0.015] = –0.054). An alternative way to contextualize these results would involve hiring one Relay teacher for every four core academic teachers. Marginal effects of this change, assuming a linear dosage effect, equate to a reduction of 0.1 absences. The magnitude of these effects aligns with findings from other studies evaluating program effects on attendance. For instance, research on student–teacher racial compatibility suggests that a racial mismatch increases the number of absences by 0.04 days per year (Holt & Gershenson, 2019).

Next, I examine whether chronic absenteeism (being absent for at least 10% of all school days) is reduced when students are taught by Relay teachers. The results in Panel B indicate that students taught by Relay teachers are 1.9 percentage points less likely to be chronically absent when controlling for classroom, school, teacher characteristics, and school fixed effects. The student fixed effects results in Column 5 indicate that the Relay program decreased the probability of experiencing chronic absences by 0.2 percentage points, equivalent to an 11% reduction from the baseline of 17.5%, which is not statistically significant. This lack of significant effect may be due to chronic absenteeism being an extreme behavior influenced by various factors beyond in-classroom interactions with Relay teachers, such as personal, familial, or socioeconomic issues.

Together, results suggest the importance of controlling for student fixed effects because they account for unobserved individual characteristics not captured by classroom, school, and teacher characteristics. This approach helps isolate the effects of the Relay program by controlling for personal invariant factors that affect attendance behaviors.

### 5.2 Relay effects on suspension outcome

Results presented in Columns 1 to 3 in Panel C indicate that the addition of classroom, school, and teacher controls reduces the Relay effects on the standardized number of suspensions from 0.058 to 0.025, which is not statistically significant. Additionally controlling for school fixed effects changes the coefficient to -0.022, which remains insignificant. Results in Column 5, which include student fixed effects, suggest that students who took all of their courses with Relay-trained teachers have 0.04 fewer standardized suspensions, compared to when they took no courses with Relay teachers, although this result is only marginally significant.

Given that one standard deviation of the number of principal suspensions is 0.210, this effect translates into a 20% reduction in the number of principal suspensions. Note again that the estimated coefficient of 0.040 represents the Relay program effects when students were taught entirely by Relay teachers (i.e., for 100% of their courses) compared to when they took no courses with Relay teachers (i.e., 0% of their courses). Since hiring Relay teachers exclusively across all NYC schools is impractical, I contextualize these results in a hypothetical scenario. By assessing the effects of recruiting one Relay teacher for one of the four main academic subjects, I provide another perspective. This adjustment translates to a reduction of 3,270 suspension cases (–0.040 × ¼ × 0.138 × 2,370,082 = –3270) among third through eighth grade students in NYC from 2014 to 2019, corresponding to a 7% decrease in the number of principal suspensions.

I also estimate Relay effects on another suspension outcome. In particular, I construct the number of any suspensions, either by the principal or superintendent. Results in Panel D show that Relay-trained teachers decrease the number of any suspensions, either principal or superintendent, by 0.03 suspensions, which is not statistically significant. Superintendent suspensions are usually issued for severe infractions like weapon possession and are less likely to be influenced by interactions with classroom teachers. Due to the severity and rarity of superintendent suspensions, I focus on the number of principal suspensions.

These suspension results are consistent with attendance outcomes, indicating that the Relay program does not affect extreme behaviors like superintendent suspensions and chronic absences. Instead, results suggest that Relay teachers are more likely to influence moderate behaviors, such as reducing the number of absences and principal suspensions, through their direct interactions with students in the classroom.

## 6. Heterogeneous Effects

The main results from Table 3 are based on the full sample. However, given existing findings on teachers' effects on student non-test outcomes, one could reasonably infer that Relay program effects differ by student characteristics. The Relay Graduate School of Education collaborates with high-needs schools, often serving students of color – especially Black students – who tend to have more absences and suspensions compared to their White peers (e.g., Gershenson, 2016). The significance of understanding this variation is underscored by the fact that the benefits of better attendance and fewer suspensions are more marked for low-income students and students of color (Gershenson et al., 2017). Thus, it is important to examine whether the effects of Relay teachers on non-test outcomes are larger among students who are more exposed to Relay teachers.

First, I explore differential program effects by race and ethnicity. Table 4 shows the estimated effects of Relay teachers on student non-test outcomes by race/ethnicity (Columns 2–3), gender (Columns 4 and 5), and for Black or Hispanic males specifically (Column 6). Column 1 presents the main results using the full sample (as reported in Column 5 in Table 3). I also present the mean of outcomes by demographic subgroup, so that I can compare the effect of Relay teachers across subgroups. For example, the mean standardized number of absences is the largest among Black and Hispanic students, indicating that these students of color have the highest number of absences (as reported in Appendix Table 4).

The estimated coefficients shown in Columns 2 and 3 in Panel A indicate Relay program effects on reducing absences for students of color. I find no Relay effects among White students, who have a relatively smaller number of absences compared to these students of color. When I examine Relay effects by gender, results show that the effects on absences are statistically significant in both groups.

Panel B presents estimated Relay effects on the standardized number of principal suspensions. Results show similar patterns, in that I find relatively large Relay effects on students of color and that I fail to find any Relay effects among White students, who have a relatively smaller number of suspensions compared to Black and Hispanic students. Columns 4 and 5 indicate that Relay effects on principal suspensions are mainly driven by male students. I do not find any significant effects among female students.

Most notably, Column 6 reveals that Relay teachers have particularly strong effects on Black or Hispanic male students. For this demographic group, Relay teachers reduce standardized absences by 0.050 standard deviations ($p<0.05$) and suspensions by 0.079 standard deviations ($p<0.05$) – nearly double the effect size observed in the overall sample. These findings are especially important given that male students of color often face disproportionate disciplinary actions in schools and higher rates of absenteeism.

To contextualize the results, I compare the magnitude from the literature on teacher–student racial matching, which shows that if a Black student has at least one Black teacher in their school and grade level, they are less likely to receive suspensions in a given year. For example, Lindsay and Hart (2017) found that if Black middle school students were exposed to all-Black teachers compared to entirely non-Black teachers, they were 1.7 percentage points less likely to get suspended in a given year. My findings suggest that having 1 Relay teacher out of 4 subjects would reduce the number of principal suspensions by 0.9 percentage points ($–0.045 \times \frac{1}{1.233} \times \frac{1}{4} = –0.009$), given that one standard deviation in principal suspensions for males is 1.233. This magnitude is comparable to the same-race teacher effects for Black students, when being exposed to half of their teachers being Black ($0.017 \times \frac{1}{2}= 0.0085$). For Black or Hispanic male students specifically, the effect would be even larger at 1.6 percentage points ($–0.079 \times \frac{1}{1.233} \times \frac{1}{4} = –0.016$).

Together these findings suggest that Relay teachers are effective in improving attendance and reducing suspensions, particularly among students of color and male students, with especially pronounced effects for male students of color. These results have important equity implications, suggesting that Relay teachers may be particularly effective at addressing persistent disparities in disciplinary practices and attendance patterns.

Second, I examine whether Relay effects differ by longitudinal student program participation status, specifically for ED. Because students who have always been ED are shown to be different from students who were eligible for only one or two years (Michelmore & Dynarski, 2017), I calculate time-invariant program variables to reflect whether students were (a) ever, (b) always, or (c) never identified as ED across years. For example, if a student was ED during all observed years, then the student is labeled as `Always-ED.' I estimate the Relay effects separately for each group.[5]

Table 5 shows that the effects of Relay teachers on the standardized number of absences in Panel A are statistically significant among students who were Always-ED (Column 3) and Ever-ED (Column 4). In contrast, Relay effects are insignificant among students who were Never-ED, or most advantaged (Column 2). This indicates that Relay effects on absences are larger among students with disadvantaged backgrounds.

Results in Panel B show that Relay teacher effects on principal suspensions are between -0.031 to -0.043, which are similar to the effects in the full sample. However, all of these estimates are insignificant. Overall, these findings suggest that Relay teachers are particularly effective in reducing absences among economically disadvantaged students, while their effects on suspension are less clear.

Finally, I estimate the effect of Relay teachers on student non-test outcomes separately by grade band, with Grades 3–5 defined as elementary school and Grades 6–8 as middle school. Table 6 summarizes the relationship between exposure to Relay teachers and student non-test outcomes for these grade bands. Column 1 includes all grades (and is thus identical to Column 5 in Table 3). Column 2 presents the results for elementary grades, and Column 3 for middle school grades. Results for number of absences (Panel A) and the number of principal suspensions (Panel B) indicate that the effects of Relay teachers vary by grade level. Although Relay teachers significantly reduce absences and marginally reduce suspensions for all grades (3rd–8th), in middle grades (6th–8th), Relay teachers show a marginally significant reduction in absences but not in suspensions. However, in elementary grades (3rd–5th), there are no significant effects on either absences or suspensions. Overall, these results suggest that Relay teachers have a more pronounced effect on reducing absenteeism and suspensions in higher grade levels, particularly middle school, while the effects are less evident in elementary grades.

## 7. Robustness

The primary specification includes the average years of teacher experience, assuming that the counterfactual for Relay teachers consists of non-Relay teachers with equivalent experience levels. However, for district or school administrators considering hiring Relay teachers, a more pertinent comparison group for the counterfactual could involve average non-Relay teachers, who are likely to have eight additional years of experience (Table 1).

In an alternative specification, I do not control for teacher experience; instead, I make the comparison between Relay teachers and the average teacher from the non-Relay distribution. Results in Panel B in Table 7 demonstrate that the coefficient estimates for Relay teachers show a slight decrease in magnitude compared to Panel A, yet they remain statistically significant for absences. However, for principal suspensions, the coefficient estimate shifts from -0.040 to -0.038, rendering it statistically insignificant. This suggests that, while Relay teachers continue to show effectiveness in reducing absences compared to more experienced non-Relay teachers, their effects on principal suspensions do not appear to differ significantly from those of their more experienced counterparts.

[5] Note that, unlike ED and ELL status, existing studies show that IEP eligibility and participation differ across schools and/or within school (e.g., Elder et al. 2021). In NYC, there are 13 disability classifications (https://www.schools.nyc.gov/learning/special-education/the-iep-process/the-iep). Thus, I do not estimate Relay effects by IEP participation status.

Next, I exclude teacher race variables. Previous sections demonstrate that Relay teachers differ from non-Relay teachers not only in terms of the students they teach but also with respect to their demographics (Table 1). If having a Relay teacher increases a student's likelihood of being taught by a teacher of color or one whose race matches theirs, the inclusion of teacher race variables could control away Relay effects operating through teacher race effects. As shown in Panel C, the main findings remain unchanged when excluding teacher race variables.

Panel D presents results from models that exclude school fixed effects. As previously discussed in the method section, there are compelling reasons to include school fixed effects in my analysis – particularly given that Relay teachers predominantly serve schools with higher populations of students of color and the potential for different disciplinary practices across schools. For absences, the estimated coefficient increases slightly in magnitude to -0.050 while maintaining statistical significance, suggesting that Relay teachers' positive effects on attendance are robust across different school environments. However, for suspensions, the coefficient decreases substantially to -0.006 and becomes statistically insignificant. This contrast between outcomes indicates that school-level factors are likely to play a critical role in explaining Relay teachers' effects on disciplinary outcomes but less so for attendance. The pattern aligns with literature suggesting that suspension practices vary considerably across schools, while attendance behaviors may be more directly influenced by classroom-level teacher-student interactions. These findings further justify the inclusion of school fixed effects in the baseline specification, particularly for accurately estimating effects on disciplinary outcomes.

Finally, I explore alternative specifications using different combinations of fixed effects. Panel E represents estimates using grade-by-year fixed effects instead of separate grade fixed effects. The results remain qualitatively similar, with coefficients maintaining similar magnitudes and statistical significance. This consistency is notable despite the fact that grade-by-year fixed effects are collinear with grade or year fixed effects within students, except for grade repeaters. These alternative specifications mainly serve as robustness checks, confirming that the main findings are not sensitive to the specific choice of fixed effects structure.

## 8. Conclusion and Discussion

This study provides the first rigorous evaluation of the Relay Graduate School of Education's impact on important student non-test outcomes. While much research finds that general teacher characteristics, such as certification type, advanced degrees, and experience, weakly or inconsistently predict student outcomes, my findings demonstrate that the Relay program meaningfully improves student attendance and reduces suspensions, with particularly large effects for students of color and male students. Using longitudinal administrative data from New York City public schools and employing student fixed-effects models to identify plausibly causal effects, I estimate the impact of Relay-trained teachers on student attendance and suspensions for students in Grades 3 through 8 between 2014 and 2019.

The findings indicate that Relay teachers are more effective than their non-Relay counterparts in improving student attendance and reducing principal suspensions. Importantly, these positive effects are particularly pronounced for historically marginalized student populations, including students of color and male students, and those with Economically Disadvantaged or English Language Learner status. Descriptive evidence also suggests that Relay teachers are more likely to be assigned to teach students who tend to experience higher rates of absenteeism and suspensions, highlighting that Relay teachers appear to be most effective with the students who may benefit most from improved engagement and behavioral support. These differential effects offer suggestive evidence that could be consistent with Relay's emphasis on culturally responsive pedagogy and comprehensive classroom management strategies, as described in the Program Description. Relay's approach may potentially equip teachers with skills and perspectives that are especially effective in creating supportive and engaging classroom environments for students from diverse backgrounds, and it is plausible that these specific elements of the training contribute to the observed differential effects for these student groups.

These results also suggest significant policy implications, particularly for districts seeking strategies to improve student engagement and reduce discipline disparities. Relay's model of embedding culturally responsive and classroom-management training directly into teacher preparation appears especially effective with historically marginalized student populations, highlighting the potential for similar models to contribute meaningfully toward addressing educational inequities.

Below, I include a few considerations for interpreting these results, as well as possible future directions for this research. First, Relay teachers are different from non-Relay teachers – not only in terms of the students they teach but also with respect to their demographics. Descriptive analyses indicate that Relay teachers are more diverse (i.e., there are more Black teachers) than non-Relay teachers. The extent to which the Relay program has improved the diversity of the teacher workforce, and how this has affected students' outcomes, is important to study. For example, as Relay teachers are more racially diverse, if students of color are more likely to be taught by first-year Relay teachers of color (especially teachers of the same race/ethnicity), the future study can explore whether this matching may be the cause of the students' improved attendance and reduced suspensions.

Second, there are many other student outcomes – beyond test scores, which are typically the easiest and most common to measure – that matter for student success. For instance, GPA and on-grade progression (i.e., on-time promotion to the next grade) are shown to be predictive of students' long-term outcomes, including high school graduation, college enrollment, and labor earnings. Exploring whether Relay teachers also have an effect on these other student outcomes could broaden our understanding of the effects that Relay teachers can have on a range of student success factors.

Third, as there are multiple pathways for teachers in NYC to enter teaching (such as traditional university-based programs and other teaching fellow programs), it is important to consider how Relay teachers' effects may compare to those of other kinds of teachers, based on factors like their educational background or training. In this paper, I group all other pathways together as non-Relay pathways. However, there may be important differences in how Relay teachers' effects compare to those of teachers from other backgrounds. Access to other teachers' pathway information would help us more precisely estimate the effectiveness of the Relay program. Fourth, while this study focused specifically on non-test outcomes due to the scope of the original research agreement, future research should examine Relay's effects on standardized test performance to provide a more comprehensive evaluation of the program's impact.

Finally, additional data on absences and suspensions are likely to improve our understanding of the effectiveness of the Relay program. As unexcused absences are shown to have more harmful effects on student achievement compared to excused absences, more detailed information on these absences (e.g., distinguishing between excused and unexcused absences and/or classroom-level absences) would be needed, to help us better understand the mechanism through which the Relay program is affecting absences. Similarly, it would also be interesting to examine which types of incidents are reduced when students are taught by Relay teachers. As the type of discipline varies from level 1 (uncooperative/noncompliant behavior, such as unexcused absences and being late for school) to level 5 (seriously dangerous or violent behavior),[6] this information would help us understand the margin along which the reduction of the suspension occurs. For example, Lindsay and Hart (2017) show that exposure to Black teachers improves student behaviors, including reducing the number of violent incidents. If a student received four days of suspension for Level 3 discipline (e.g., vandalism) in the previous year, taking no courses with a Relay teacher, but receives four days of suspension due to Level 2 discipline (e.g., possession of cigarettes) when taught only by Relay teachers, neither number of suspensions nor number of days suspended can capture the improved behavior.

## Acknowledgements

An earlier version of this study was commissioned by Relay Graduate School of Education and Education Analytics and has been available as a working paper on SSRN since May 2023 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4437634). The current manuscript represents independent academic research conducted since then, with substantial revisions to the methodology, analysis, and presentation.

I thank my former colleagues at Education Analytics for their contributions during the early stages of this project: Robert Meyer for his guidance during project development and for enabling my continued access to the data after my departure from Education Analytics; Joshua Marland, Nick Masada, and Ernest Morgan for project coordination and partnership management; and Abhijeet Ball and Annie Nielsen for research assistance. I am grateful to Anne-Marie Hoxie, Andrew Lumpe, Jung-In Soh, and Amber Wang at Relay Graduate School of Education for their support of this research direction and throughout the data collection process. I particularly thank Anthony Milanowski for his thoughtful feedback on manuscript revisions. I also benefited from helpful comments by seminar participants at the Association for Education Finance and Policy, the Society for Research on Educational Effectiveness, and the University of Wisconsin–Madison.

The views expressed in this paper are solely those of the author and do not represent the positions of any previously affiliated organizations.

[6] For K–5 students, see https://www.schools.nyc.gov/docs/default-source/default-document-library/discipline-code-kindergarten-grade-5-english. For 6th to 12th graders, see https://www.schools.nyc.gov/docs/default-source/default-document-library/discipline-code-grade-6-12-english.

# References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics*, *138*. https://doi.org/10.1093/qje/qjac038

An, Y., & Koedel, C. (2021). How do teachers from alternative pathways contribute to the teaching workforce in urban areas? Evidence from Kansas City. *AERA Open*, *7*, 23328584211026952.

Antecol, H., Eren, O., & Ozbeklik, S. (2013). The effect of Teach for America on the distribution of student achievement in primary school: Evidence from a randomized experiment. *Economics of Education Review*, *37*, 113-125.

Aucejo, E. M., & Romano, T. F. (2016). Assessing the effect of school days and absences on test score performance. *Economics of Education Review*, *55*, 70–87.

Bacher-Hicks, A., Billings, S. B., & Deming, D. J. (2024). The School-to-Prison Pipeline: Long-Run Impacts of School Suspensions on Adult Crime. *American Economic Journal: Economic Policy*, *16*(4), 165-193.

Backes, B., & Hansen, M. (2018). The impact of Teach for America on non-test academic outcomes. *Education Finance and Policy*, *13*(2), 168–193.

Backes, B., & Hansen, M. (2024). Estimates of Teach for America Effects on Student Test and Nontest Academic Outcomes Over Time. *AERA Open*, *10*, 23328584241234874.

Blazar, D. (2021). *Teachers of color, culturally responsive teaching, and student outcomes: Experimental evidence from the random assignment of teachers to classes* (EdWorkingPaper No. 21-501). Annenberg Institute, Brown University. https://doi.org/10.26300/jym0-wz02

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

Bowes, N-K. E. (2017). *Culturally responsive academic advisement*. Doctoral Dissertation, Johns Hopkins University, Baltimore, MD.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440.

Clark, M. A., & Isenberg, E. (2020). Do Teach For America Corps Members Still Improve Student Achievement? Evidence from a Randomized Controlled Trial of Teach For America's Scale-Up Effort. *Education Finance and Policy*, *15*(4), 736-760.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of education review*, *26*(6), 673-682.

Cochran-Smith, M., & Reagan, E. M. (2021). *"Best practices" for evaluating teacher preparation programs*. National Academy of Education Committee on Evaluating and Improving Teacher Preparation Programs.

Curtis, A. J. (2013). Tracing the school-to-prison pipeline from zero-tolerance policies to juvenile justice dispositions. *Georgetown Law Journal*, *102*(4), 1251.

Davison, M., Penner, A. M., Penner, E., Pharris-Ciurej, N., Porter, S. R., Rose, E., Shem-Tov, Y., & Yoo, P. (2022). School discipline and racial disparities in early adulthood. *Educational Researcher*, *51*(3), 231–234.

Elder, T. E., Figlio, D. N., Imberman, S. A., & Persico, C. L. (2021). School segregation and racial gaps in special education identification. *Journal of Labor Economics*, *39*(S1), S151–S197.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*(2), 125–149.

Gershenson, S., Jacknowitz, A., & Brannegan, A. (2017). Are student absences worth the worry in US primary schools? *Education Finance and Policy*, *12*(2), 137–165.

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

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational evaluation and policy analysis*, *22*(2), 129-145.

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29–44.

Gottfried, M A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 392–419.

Gottfried, M A. (2011). The detrimental effects of missing school: Evidence from urban siblings. *American Journal of Education*, *117*(2), 147–182.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*(3), 411–482.

Henry, G. T., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Purtell, K. M., Thompson, C. L., & Zulli, R. A. (2014). Teacher preparation policies and their effects on student achievement. *Education Finance and Policy*, *9*(3), 264-303.

Holt, S. B., & Gershenson, S. (2019). The impact of demographic representation on absences and suspensions. *Policy Studies Journal*, *47*(4), 1069–1099.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education review*, *27*(6), 615-631.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher preparation programs and teacher quality: Are there real differences across programs? *Education Finance and Policy*, *10*(4), 508–534.

Lindsay, C. A., & Hart, C. M. (2017). Exposure to same-race teachers and student disciplinary outcomes for Black students in North Carolina. *Educational Evaluation and Policy Analysis*, *39*(3), 485–510.

Liu, J., Lee, M., & Gershenson, S. (2021). The short-and long-run impacts of secondary school absences. *Journal of Public Economics*, *199*, 104441.

Lovison, V. S. (2025). The effects of high-performing, high-turnover teachers on long-run student achievement: Evidence from Teach For America. *Educational Evaluation and Policy Analysis*, *47*(2), 636-642.

Michelmore, K., & Dynarski, S. (2017). The gap within the gap: Using longitudinal data to understand income differences in educational outcomes. *AERA Open*, *3*(1), 2332858417692958.

Morris, E. W., & Perry, B. L. (2016). The punishment gap: School suspension and racial disparities in achievement. *Social Problems*, *63*(1), 68–86.

Mungal, A. S. (2016). Teach For America, Relay Graduate School, and the charter school networks: The making of a parallel education structure. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas*, *24*, 1-30.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

Pearman, F. A., Curran, F. C., Fisher, B., & Gardella, J. (2019). Are achievement gaps related to discipline gaps? Evidence from national data. *Aera Open*, *5*(4), 2332858419875440.

Penner, E. K. (2016). Teaching for all? Teach for America's effects across the distribution of student achievement. *Journal of Research on Educational Effectiveness*, *9*(3), 259-282.

Penner, E. K. (2021). Teach For America and teacher quality: Increasing achievement over time. *Educational Policy*, *35*(7), 1047-1084.

Tran, L., & Gershenson, S. (2021). Experimental estimates of the student attendance production function. *Educational Evaluation and Policy Analysis*, *43*(2), 183–199.

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? *Economics of Education Review*, *53*, 31–45.

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach for America in high school. *Journal of Policy Analysis and Management*, *30*(3), 447–469.

Table 1. Descriptive Statistics

| | All<br>(1) | Relay<br>(2) | Non-relay<br>(3) | Mean<br>difference<br>(2–3)<br>(4) |
|---|---|---|---|---|
| Panel A: Teachers | | | | |
| Female | 0.787 | 0.724 | 0.788 | -0.064$^*$ |
| White | 0.583 | 0.460 | 0.584 | -0.124$^*$ |
| Black | 0.189 | 0.242 | 0.189 | 0.053$^*$ |
| Hispanic | 0.147 | 0.164 | 0.147 | 0.018 |
| Asian | 0.056 | 0.104 | 0.056 | 0.049$^*$ |
| Mean years of experience | 9.893 | 1.259 | 9.893 | -8.634$^*$ |
| Number of teachers | 151,964 | 1,083 | 150,881 | |
| Panel B: Student demographics | | | | |
| Female | 0.485 | 0.485 | 0.485 | 0.001 |
| White | 0.168 | 0.073 | 0.172 | -0.099$^*$ |
| Black | 0.227 | 0.285 | 0.225 | 0.06$^*$ |
| Hispanic | 0.414 | 0.536 | 0.409 | 0.127$^*$ |
| Asian | 0.175 | 0.094 | 0.179 | -0.085$^*$ |
| Economically disadvantaged (ED) | 0.727 | 0.837 | 0.722 | 0.115$^*$ |
| Individualized Education Plan (IEP) | 0.223 | 0.246 | 0.223 | 0.023$^*$ |
| English Language Learner (ELL) | 0.12 | 0.128 | 0.12 | 0.008$^*$ |
| Number of students | 2,263,127 | 86,534 | 2,176,593 | |
| Panel C: Student non-test outcomes | | | | |
| Number of absences | 10.216 | 12.096 | 10.142 | 1.954$^*$ |
| Chronic absence | 0.16 | 0.199 | 0.159 | 0.04$^*$ |
| Number of principal suspensions | 0.025 | 0.049 | 0.024 | 0.026$^*$ |
| Number of superintendent suspensions | 0.01 | 0.023 | 0.01 | 0.014$^*$ |
| Received any principal suspensions | 0.019 | 0.037 | 0.018 | 0.018$^*$ |
| Received any superintendent suspensions | 0.008 | 0.019 | 0.008 | 0.011$^*$ |
| Panel D: Student prior outcomes | | | | |
| Math test scores | 0.024 | -0.26 | 0.038 | -0.298$^*$ |
| ELA test scores | 0.022 | -0.237 | 0.034 | -0.271$^*$ |
| Number of absences | 9.932 | 11.324 | 9.877 | 1.446$^*$ |
| Number of principal suspensions | 0.019 | 0.038 | 0.019 | 0.019$^*$ |
| Number of superintendent suspensions | 0.006 | 0.014 | 0.006 | 0.008$^*$ |
| Received any principal suspensions | 0.014 | 0.028 | 0.014 | 0.014$^*$ |
| Received any superintendent suspensions | 0.005 | 0.012 | 0.005 | 0.007$^*$ |

*Note*. The analytic sample consists of students in Grades 3–8 between 2014 and 2019 and teachers in Grades 3–8 between 2014 and 2018. `Relay students' refers to students who took at least one course from Relay teachers. `Non-Relay students' refers to students who took no courses with Relay teachers. Column 4 shows that the difference in means between Relay students and non-Relay students is statistically significant at $p < 0.05$.

Table 2. Non-Random Sorting of Students to Relay Teachers

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Black | 0.008$^{***}$ | 0.001 | 0.001 | 0.001 | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Hispanic | 0.008$^{***}$ | 0.001 | 0.001 | 0.001 | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Asian | -0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| ELL | 0.001 | -0.002$^{***}$ | -0.002$^{***}$ | -0.002$^{***}$ | -0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| IEP | -0.001$^{**}$ | -0.001$^{***}$ | -0.001$^{***}$ | -0.001$^{***}$ | -0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| ED | 0.003$^{***}$ | 0.001 | 0.001 | 0.001 | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Female | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Obs. | 2,887,735 | 2,823,663 | 2,823,662 | 2,823,662 | 2,601,648 |
| School controls | | Yes | Yes | Yes | Yes |
| Classroom controls | | | Yes | Yes | Yes |
| School FE | | | | Yes | Yes |
| Student FE | | | | | Yes |

*Note*. The dependent variable is the Relay share dosage variable. Each column represents a different regression specification: (1) Base model with student demographics and grade fixed effects; (2) Adds school-level controls including average test scores and demographic compositions; (3) Adds classroom-level peer controls; (4) Adds school fixed effects; (5) Adds student fixed effects. Student demographics include ED, IEP, ELL status, gender, and race/ethnicity. School-level controls include average lagged test scores and proportions of student demographic characteristics. Classroom-level controls include average lagged test scores and demographic composition of peers. Standard errors clustered at the school level are in parentheses.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table 3. Estimated Relay Program Effects on Student Outcomes

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Panel A. Standardized number of absences | | | | | |
| Relay | 0.084 | -0.015 | -0.047$^{*}$ | -0.038$^{*}$ | -0.040$^{**}$ |
| | (0.057) | (0.028) | (0.028) | (0.022) | (0.018) |
| Panel B. Chronic absences | | | | | |
| Relay | 0.034$^{*}$ | 0.001 | -0.014 | -0.019$^{***}$ | -0.002 |
| | (0.020) | (0.010) | (0.009) | (0.007) | (0.007) |
| Panel C. Standardized number of principal suspensions | | | | | |
| Relay | 0.058 | 0.057 | 0.025 | -0.022 | -0.040$^{*}$ |
| | (0.046) | (0.044) | (0.042) | (0.026) | (0.024) |
| Panel D. Standardized number of any suspensions | | | | | |
| Relay | 0.079$^{*}$ | 0.066$^{*}$ | 0.033 | -0.017 | -0.026 |
| | (0.043) | (0.040) | (0.038) | (0.024) | (0.023) |
| Observations | 2,604,457 | 2,598,140 | 2,597,512 | 2,597,492 | 2,370,082 |
| School controls | | Yes | Yes | Yes | Yes |
| Classroom controls | | | Yes | Yes | Yes |
| Teacher controls | | | Yes | Yes | Yes |
| School fixed effects | | | | Yes | Yes |
| Student fixed effects | | | | | Yes |

*Note*. The sample is restricted to students in Grades 3–8 between 2014 and 2019. The outcomes are the number of absences standardized by grade and by year in Panel A, the likelihood of being chronically absent in Panel B, the number of principal suspensions standardized by grade and by year in Panel C, and number of principal and superintendent suspensions standardized by grade and by year in Panel D. Each column includes controls for student demographics (ED, IEP, and ELL) and grade fixed effects. Classroom-level controls include average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL. School-level controls include average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL. Teacher-level controls include proportions of teacher gender and race/ethnicity and average years of experience. Clustered standard errors are in parentheses.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table 4. Relay Effects on Non-Test Outcomes, by Student's Race/Ethnicity and Gender

| | All | Black or Hispanic | White | Male | Female | Male Black or Hispanic |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Panel A: Standardized number of absences | | | | | | |
| Relay | -0.040$^{**}$ | -0.045$^{**}$ | -0.017 | -0.045$^{**}$ | -0.034$^{**}$ | -0.050$^{**}$ |
| | (0.018) | (0.020) | (0.044) | (0.023) | (0.020) | (0.025) |
| $R^2$ | 0.737 | 0.723 | 0.737 | 0.740 | 0.735 | 0.727 |
| Mean | -0.002 | 0.141 | -0.123 | 0.014 | -0.031 | 0.166 |
| | | | | | | |
| Panel B: Standardized number of suspensions | | | | | | |
| Relay | -0.040$^{*}$ | -0.055$^{**}$ | 0.046 | -0.070$^{**}$ | -0.005 | -0.079$^{**}$ |
| | (0.024) | (0.026) | (0.082) | (0.032) | (0.029) | (0.035) |
| $R^2$ | 0.398 | 0.397 | 0.413 | 0.405 | 0.372 | 0.405 |
| Mean | -0.002 | 0.027 | -0.040 | 0.050 | -0.056 | 0.087 |
| | | | | | | |
| Observations | 2,370,732 | 1,541,780 | 379,667 | 1,220,595 | 1,149,470 | 791,683 |

*Note*. The sample is restricted to students in Grades 3–8 between 2014 and 2019. The outcomes are the number of absences standardized by grade and by year in Panel A and the number of principal suspensions standardized by grade and by year in Panel B. Each column includes controls for student demographics (ED, IEP, and ELL), school-level controls (average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL), teacher-level controls (proportions of teacher gender and race/ethnicity and average years of experience), and classroom-level controls (average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL), grade fixed effects, student fixed effects, and school fixed effects. Clustered standard errors are in parentheses.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table 5. Relay Effects on Non-Test Outcomes, by Student's Longitudinal Program Status

| | All<br>(1) | Never ED<br>(2) | Always ED<br>(3) | Ever ED<br>(4) |
|---|---|---|---|---|
| Panel A: Standardized number of absences | | | | |
| Relay | -0.040$^{**}$ | -0.009 | -0.036$^{**}$ | -0.056$^{***}$ |
| | (0.018) | (0.034) | (0.021) | (0.021) |
| $R^2$ | 0.737 | 0.727 | 0.727 | 0.739 |
| Mean | -0.002 | -0.311 | 0.128 | -0.087 |
| | | | | |
| Panel B: Standardized number of principal suspensions | | | | |
| Relay | -0.040$^{**}$ | -0.039 | -0.043 | -0.031 |
| | (0.024) | (0.037) | (0.029) | (0.031) |
| $R^2$ | 0.398 | 0.427 | 0.397 | 0.398 |
| Mean | -0.002 | -0.067 | 0.022 | -0.013 |
| | | 35% | 91% | 79% |
| Observations | 2,370,732 | 301,665 | 1,251,882 | 816,482 |

*Note*. The sample is restricted to students in Grades 3–8 between 2014 and 2019. The outcomes are the number of absences standardized by grade and by year in Panel A and the number of principal suspensions standardized by grade and by year in Panel B. Each column includes controls for student demographics (ED, IEP, and ELL), school-level controls (average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL), teacher-level controls (proportions of teacher gender and race/ethnicity and average years of experience), and classroom-level controls (average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL), grade fixed effects, student fixed effects, and school fixed effects. Clustered standard errors are in parentheses.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table 6. Relay Effects on Non-Test Outcomes, by Grade Bands

| | All Grades | Elementary Grades | Middle Grades |
|---|---|---|---|
| | (3rd–8th) | (3rd–5th) | (6th–8th) |
| | (1) | (2) | (3) |
| Panel A: Standardized number of absences | | | |
| Relay | -0.040** | -0.025 | -0.044** |
| | (0.018) | (0.038) | (0.026) |
| $R^2$ | 0.737 | 0.798 | 0.791 |
| Mean | -0.007 | 0.003 | -0.023 |
| | | | |
| Panel B: Standardized number of principal suspensions | | | |
| Relay | -0.040** | -0.002 | -0.035 |
| | (0.024) | (0.071) | (0.033) |
| $R^2$ | 0.398 | 0.466 | 0.533 |
| Mean | -0.002 | -0.001 | -0.033 |
| Observations | 2,370,082 | 1,141,487 | 1,064,101 |

*Note*. The sample is restricted to students in Grades 3–8 between 2014 and 2019 in Column 1. Column 2 restricts the sample to students in Grades 3–5, and Column 3 restricts the sample to Grades 6 to 8. The outcome is the number of absences standardized by grade and by year in Panel A and the number of principal suspensions standardized by grade and by year in Panel B. Each column includes controls for student demographics (ED, IEP, and ELL), school-level controls (average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL), teacher-level controls (proportions of teacher gender and race/ethnicity and average years of experience), and classroom-level controls (average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL), grade fixed effects, student fixed effects, and school fixed effects. Clustered standard errors are in parentheses.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Table 7. Robustness of Estimates

| | Standardized number of absences | Standardized number of principal suspensions |
|---|---|---|
| | (1) | (2) |
| Panel A: Baseline | | |
| Relay | -0.040$^{**}$ | -0.040$^{**}$ |
| | (0.018) | (0.024) |
| $R^2$ | 0.737 | 0.398 |
| | | |
| Panel B: Exclude teacher experience variables | | |
| Relay | -0.039$^{**}$ | -0.038 |
| | (0.018) | (0.024) |
| $R^2$ | 0.737 | 0.398 |
| | | |
| Panel C: Exclude teacher race variables | | |
| Relay | -0.037$^{**}$ | -0.040$^{**}$ |
| | (0.018) | (0.024) |
| $R^2$ | 0.737 | 0.398 |
| | | |
| Panel D: Without school fixed effects | | |
| Relay | 0.050$^{***}$ | -0.006 |
| | (0.022) | (0.026) |
| $R^2$ | 0.734 | 0.392 |
| | | |
| Panel E: Grady-by-year fixed effects | | |
| Relay | 0.048$^{***}$ | -0.041$^{*}$ |
| | (0.018) | (0.024) |
| $R^2$ | 0.738 | 0.398 |
| Observations | 2,370,082 | 2,370,082 |

*Note*. Each column includes controls for student demographics (ED, IEP, and ELL), grade fixed effects, classroom-level controls (average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL), school-level controls (average of lagged test scores, proportions of student gender, race/ethnicity, ED, IEP, and ELL), teacher-level controls (proportions of teacher gender and race/ethnicity and average years of experience), and school fixed effects. Clustered standard errors are in parentheses.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Appendix Table 1. Total Number of Relay and Non-Relay Teachers in New York City Public Schools, by Year, Overall, and Among First-Year Teachers

| | First-Year Teachers | | Overall | |
|---|---|---|---|---|
| | Relay | Non-Relay | Relay | Non-Relay |
| Year | (1) | (2) | (3) | (4) |
| 2014 | 64 | 2,176 | 112 | 27,644 |
| 2015 | 66 | 2,639 | 151 | 29,307 |
| 2016 | 61 | 2,583 | 182 | 30,344 |
| 2017 | 101 | 2,733 | 251 | 31,560 |
| 2018 | 167 | 2,483 | 387 | 32,026 |
| 2019 | 139 | 2,176 | 465 | 32,062 |
| Total | 598 | 14,790 | 1,548 | 182,943 |

*Note*. The sample is restricted to teachers who taught any courses in Grades 3–8 in NYC between 2014–2019. Since the teacher demographics information for 2019 are unavailable, Table 1 uses the teacher sample data from 2014 to 2018. However, this table includes all counts from 2014 to 2019.

Appendix Table 2. Descriptive Statistics of Relay and Non-Relay Teachers in New York City Public Schools, Among First-Year Teachers in Grades 3–8

| | All<br>(1) | Relay<br>(2) | Non-Relay<br>(3) | Mean difference<br>(2–3)<br>(4) |
|---|---|---|---|---|
| Female | 0.789 | 0.706 | 0.792 | -0.086* |
| White | 0.576 | 0.438 | 0.582 | -0.144* |
| Black | 0.152 | 0.258 | 0.147 | 0.111* |
| Hispanic | 0.158 | 0.164 | 0.158 | 0.006* |
| Asian | 0.08 | 0.103 | 0.079 | 0.025* |
| Mean years of experience | 0.228 | 0.183 | 0.23 | -0.047* |
| Number of teachers | 151,964 | 1,083 | 150,881 | |

*Note*. The sample is restricted to first-year teachers who taught any of the four key academic courses (math, ELA, social studies, and science) in Grades 3–8 between 2014 and 2019. It is possible for a teacher to instruct multiple courses across grades within a year. To simplify the analysis, I constructed a variable which indicates the grade in which a teacher instructed the highest proportion of their students during a school year. I use that variable to determine which grade the teacher is teaching in the analyses. *Difference in means between Relay and non-Relay teachers is statistically significant at $p < 0.05$.

Appendix Table 3. Student Characteristics Predicting Assignment to Relay Teachers, Among First-Year Teachers in Grades 3–8

| | (1) |
|---|---|
| Female | -0.001$^{**}$ |
| | (0.000) |
| Black | 0.010$^{***}$ |
| | (0.001) |
| Hispanic | 0.005$^{***}$ |
| | (0.001) |
| Asian | 0.002$^{***}$ |
| | (0.001) |
| Economically disadvantaged (ED) | 0.004$^{***}$ |
| | (0.000) |
| Individualized Education Plan (IEP) | -0.001$^{***}$ |
| | (0.000) |
| English Language Learner (ELL) | -0.014$^{***}$ |
| | (0.000) |
| Number of courses | 1,184,874 |

*Note*. The sample is restricted to courses taught by first-year teachers who taught any of the four key academic courses (math, ELA, social studies, and science) in Grades 3–8 between 2014 and 2019. Each column includes controls for student demographics (gender, race/ethnicity, ED, IEP, and ELL), grade fixed effects, year fixed effects, and district fixed effects. The White student group is the omitted reference group.

$^{*}$ $p < 0.1$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

Appendix Table 4. Differences in Student Absences

| | Mean<br>(1) | *SD*<br>(2) | *N*<br>(3) | Mean difference<br>(4) |
|---|---|---|---|---|
| Gender | | | | |
| Female | 9.135 | 8.897 | 1,078,691 | |
| Male | 9.582 | 9.168 | 1,139,723 | 0.448$^*$ |
| Race/ethnicity | | | | |
| Non-Hispanic White | 8.431 | 7.528 | 376,732 | |
| Non-Hispanic Black | 11.102 | 9.852 | 498,376 | 2.671$^*$ |
| Hispanic | 10.762 | 9.398 | 913,394 | 2.331$^*$ |
| Asian | 4.898 | 6.543 | 394,853 | -3.534$^*$ |
| Other | 9.513 | 8.862 | 17,318 | 1.082$^*$ |
| Economically disadvantaged (ED) | | | | |
| ED | 10.093 | 9.552 | 1,604,379 | 2.631$^*$ |
| Non-ED | 7.462 | 7.199 | 614,035 | |
| Always-ED | 10.623 | 9.878 | 1,131,664 | 4.168$^*$ |
| Ever-ED | 8.705 | 8.401 | 772,668 | 2.251 |
| Never-ED | 6.454 | 6.064 | 314,082 | |
| Individualized Education Plan (IEP) | | | | |
| IEP | 12.460 | 10.337 | 485,244 | 3.962$^*$ |
| Non-IEP | 8.498 | 8.441 | 1,733,170 | |
| Always-IEP | 12.739 | 10.438 | 296,773 | 4.277$^*$ |
| Ever-IEP | 11.229 | 9.886 | 264,927 | 2.767$^*$ |
| Never-IEP | 8.462 | 8.419 | 1,656,714 | |
| English Language Learner (ELL) | | | | |
| ELL | 9.402 | 8.983 | 265,982 | 0.043$^*$ |
| Non-ELL | 9.360 | 9.048 | 1,952,432 | |
| Always-ELL | 10.397 | 9.451 | 160,730 | 0.762$^*$ |
| Ever-ELL | 7.256 | 7.846 | 303,683 | -2.379$^*$ |
| Never-ELL | 9.635 | 9.142 | 1,754,001 | |
| Grade | | | | |
| Grade 3 | 9.562 | 8.871 | 384,473 | |
| Grade 4 | 9.139 | 8.860 | 385,870 | -0.423$^*$ |
| Grade 5 | 9.126 | 8.775 | 380,236 | -0.436$^*$ |
| Grade 6 | 8.904 | 8.907 | 357,870 | -0.658$^*$ |
| Grade 7 | 9.068 | 9.246 | 355,751 | -0.493$^*$ |
| Grade 8 | 10.417 | 9.521 | 354,214 | 0.855$^*$ |

*Note*. The analysis sample consists of students in Grades 3–8 between 2014 and 2019. *SD* indicates standard deviation. Mean difference *t* tests were performed to compare third to upper grades, males and females, students who are and are not ELL, students who are and are not ED, and students receiving and not receiving IEP. For longitudinal program participation status, I use the never-participated students (i.e., Never-ED, Never-IEP, and Never-ELL) as the reference group.

* Difference in means statistically significant at $p < 0.05$

Appendix Table 5. Differences in Number of Principal Suspensions

| | Mean<br>(1) | *SD*<br>(2) | *N*<br>(3) | Mean difference<br>(4) |
|---|---|---|---|---|
| Gender | | | | |
| Female | 0.005 | 0.083 | 1,081,517 | |
| Male | 0.012 | 0.129 | 1,144,598 | -0.007* |
| Race/ethnicity | | | | |
| Non-Hispanic White | 0.003 | 0.058 | 377,210 | |
| Non-Hispanic Black | 0.02 | 0.17 | 501,071 | 0.017* |
| Hispanic | 0.007 | 0.099 | 917,317 | 0.005* |
| Asian | 0.002 | 0.053 | 395,365 | -0.001* |
| Other | 0.006 | 0.093 | 17,382 | 0.004* |
| Economically disadvantaged (ED) | | | | |
| ED | 0.01 | 0.118 | 1,611,059 | 0.006* |
| Non-ED | 0.004 | 0.079 | 615,056 | |
| Always-ED | 0.011 | 0.124 | 1,136,824 | 0.009* |
| Ever-ED | 0.007 | 0.103 | 774,972 | 0.006* |
| Never-ED | 0.002 | 0.047 | 314,319 | |
| Individualized Education Plan (IEP) | | | | |
| IEP | 0.017 | 0.161 | 487,561 | 0.011* |
| Non-IEP | 0.006 | 0.089 | 1,738,554 | |
| Always-IEP | 0.016 | 0.152 | 298,203 | 0.01* |
| Ever-IEP | 0.017 | 0.163 | 267,341 | 0.011* |
| Never-IEP | 0.006 | 0.086 | 1,660,571 | |
| English Language Learner (ELL) | | | | |
| ELL | 0.006 | 0.091 | 267,936 | -0.002* |
| Non-ELL | 0.009 | 0.111 | 1,958,179 | |
| Always-ELL | 0.008 | 0.106 | 162,026 | -0.001* |
| Ever-ELL | 0.003 | 0.065 | 304,612 | -0.006* |
| Never-ELL | 0.009 | 0.115 | 1,759,477 | |
| Grade | | | | |
| Grade 3 | 0.002 | 0.051 | 387,347 | |
| Grade 4 | 0.003 | 0.07 | 387,803 | 0.002* |
| Grade 5 | 0.004 | 0.074 | 380,977 | 0.003* |
| Grade 6 | 0.009 | 0.114 | 358,780 | 0.008* |
| Grade 7 | 0.015 | 0.148 | 356,521 | 0.014* |
| Grade 8 | 0.018 | 0.157 | 354,687 | 0.016* |

*Note*. The analysis sample consists of students in Grades 3–8 between 2014 and 2019. *SD* indicates standard deviation. Mean difference *t* tests were performed to compare third to upper grades, males and females, students who are and are not ELL, students who are and are not ED, and students receiving and not receiving IEP. For longitudinal program participation status, I use the never-participated students (i.e., Never-ED, Never-IEP, and Never-ELL) as the reference group.

* Difference in means statistically significant at $p < 0.05$