Evidence on Team-Based Teaching Models in Arizona

Mary Laski
Calen Clifton
Dan Goldhaber

August 2026

WORKING PAPER No. 343-0826

CALDER
National Center for Analysis of Longitudinal Data in Education Research

AIR®

Evidence on Team-Based Teaching Models in Arizona

Mary Laski
Arizona State University

Calen Clifton
Arizona State University

Dan Goldhaber
American Institutes for Research / CALDER
University of Washington / CEDR

Contents

Acknowledgments

Support for this work was provided by the Overdeck Family Foundation and Tandem Philanthropies. Laski (mary.laski@asu.edu) is the corresponding author. We thank Ben Backes and conference participants at SREE and AEFP for helpful comments and feedback.

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

Evidence on Team-Based Teaching Models in Arizona

Mary Laski, Calen Clifton, & Dan Goldhaber

CALDER Working Paper No. 343-0826
August 2026

Abstract

Strategic staffing models are scaling rapidly nationwide. We study the impacts of a team teaching model implemented in one large district in Arizona over several years. We exploit the staggered rollout over time in a difference-in-differences design to estimate effects on student test achievement. We find overall null impacts on student learning, but important heterogeneity by grade and subject. In particular, we find consistently positive effects on middle school math achievement across all specifications, but negative effects at the elementary level in some specifications. Across grades and subjects, teamed teachers also have significantly lower turnover rates and higher evaluation ratings, but the patterns of effects are not consistent with the achievement results. The student achievement findings are broadly consistent with the literature on teacher-level departmentalization and suggest that benefits to team teaching can emerge in contexts where teaming increases collaboration without disrupting established teacher-student relationships, such as in already-departmentalized middle school grades.

1. Introduction

The traditional model of American schooling has remained largely unchanged for over a century. Critics of this "egg-crate" model note that it places an immense burden on teachers who are often isolated and have limited opportunities for specialization or collaboration (Lortie, 1975). Arguably because of this, attracting, developing, and retaining effective teachers in the United States is a long-standing challenge (Kraft & Lyon, 2024).

Many schools and districts have responded to the egg-crate critique by adopting "strategic staffing" models designed to make teaching more sustainable and effective. Policy interest in strategic staffing has accelerated in recent years and the rapid expansion of these models has led to recognition from the U.S. Department of Education, which recently authorized the use of Title II funds to support strategic staffing (U.S. Department of Education, 2026).

This paper provides the first plausibly causal evidence on one prominent strategic staffing reform: the Next Education Workforce™ (NEW), developed at Arizona State University (ASU) and now implemented in 45 school systems in 17 states. Educators in NEW models work in teams, distribute roles, and share responsibility for a single larger roster of students. We evaluate this intervention as implemented in a large, diverse school district in Arizona. We exploit the staggered, multi-year rollout of NEW across schools and grades, employing a difference-in-differences framework to estimate the effect of implementing this team-based model on student and teacher outcomes.

We find overall null impacts on student learning, measured by state assessments in math and English language arts, but also important heterogeneity by grade and subject. Elementary effects are small and negative, and often sensitive to specification decisions. We find positive, moderately large middle school effects on math tests, on the order of 0.08 SD. Middle school math estimates are consistent across specifications, and tests of pre-trends suggest these estimates are not biased. Math effects are also of similar magnitude and marginally significant in alternate models using teacher fixed effects. NEW teachers also have significantly lower turnover rates and higher evaluation ratings than comparison teachers, but the patterns of teacher effects cannot explain the gains in middle school math. Rather, exploratory analyses suggest instructional differentiation and team composition may explain the patterns of effects, though evidence on mechanisms is limited.

Following recent advances in the two-way fixed effects literature, we confirm that our estimates for middle school math are not meaningfully biased by within-treatment comparisons and remain consistent when restricted to treated and never-treated grades. They are less consistent, however, at the elementary level and for ELA. Our primary results are also robust to alternative sample restrictions and definitions of teaming.

2. Background Literature and NEW

As school systems seek to better support and deploy their workforce, strategic staffing models have emerged as a framework for rethinking how schools organize educators' roles, time, and compensation to better align educator skills with student needs (Laski et al., 2026). While there are no specific figures on the numbers of schools or districts engaged in strategic staffing initiatives, there is evidence that such initiatives have gained momentum in recent years. For instance, the federal government recently authorized Title II funds to support strategic staffing (U.S. Department of Education, 2026), and two of the largest strategic staffing models claim to have reached over 300,000 students and roughly 12,000 educators across multiple states in the 2026 school year.[1,2]

[1] Authors' calculations from information posted publicly on websites for Opportunity Culture (https://www.opportunityculture.org/opportunity-culture-stats/) and Next Education Workforce (https://workforce.education.asu.edu/).

[2] Throughout the paper, we denote each school year by its spring semester.

The intervention we focus on is "team-based" staffing based on the design of the NEW model. In contrast to traditional "one-teacher, one-classroom" models, teachers in NEW models work together in teams of roughly two to six educators, sharing responsibility for one larger roster of students. According to NEW's theory of action (Figure 1), this approach allows educators to differentiate roles and distribute responsibilities according to each member's expertise, easing the burden on each individual teacher while allowing for increased collaboration and support within the teaching team. Teams generally feature licensed teachers and paraeducators but may also include other school staff or community volunteers. In theory, NEW's approach enhances working conditions for teachers, potentially altering instructional practice and positively influencing student outcomes (Next Education Workforce, n.d.). NEW has expanded rapidly since its initial pilots in 2020; NEW teams are currently operating in 45 school systems in 17 states.

Another prominent example under the strategic staffing umbrella is Public Impact's Opportunity Culture (OC), which focuses on role differentiation and "extending the reach" of highly effective teachers. In OC models, these teachers are paid a stipend to serve as Multi-Classroom Leaders® (MCLs), leading and coaching educator teams while influencing the instructional experiences of more students. OC models may also include other formal roles that receive additional pay, such as Team Reach Teachers™ or Reach Associate™ roles (Public Impact, n.d.). OC staffing models have demonstrated impact: a quasi-experimental analysis of the model found that students taught by coached teachers scored about 0.11 SD higher in math (Backes & Hansen, 2018). More recent descriptive analyses found consistent gains in both math and reading and evidence of potential spillover effects to non-participating classrooms (Kirksey et al., 2026).

While operating in the same strategic staffing space, OC and NEW are distinct. Unlike OC, NEW does not require additional stipends to participating teachers or have a paid formal hierarchy. NEW also does not have rigid or predetermined implementation elements: principals have leeway in assigning teachers and grades to teams, and schools and teams are encouraged to adapt NEW's framework to their local context. This creates wide variation in the teamed experience.

Another common departure from the traditional school model is subject specialization or departmentalization. The evidence on specialization is mixed. Several recent quasi-experimental studies demonstrate that teachers in elementary schools see declines in value-added when they shift to a specialized model (Bastian & Fortner, 2020; Hwang & Kisida, 2022; Backes et al., 2026). The mechanisms are still poorly understood, though one hypothesis is that specialized teachers have weaker connections with their students and are less able to adapt to different student needs (Fryer, 2018).

Interestingly, despite the demonstrated negative impacts on *teacher performance*, the effect of specialization on *student achievement* is inconsistent across studies. Fryer (2018) finds significant negative impacts in Texas, while Backes et al. (2025) find positive effects that fade out in math and ELA, and Bastian and Fortner (2020) and Hwang and Kisida (2022) find null effects. The differences across outcomes are likely due to teacher sorting: even as a teacher's individual value-added declines, if the highest-performing teachers are sorted into their most impactful specialized roles, student achievement can still improve (Backes et al., 2026). While specialization is related to NEW, as NEW educators may be able to more effectively specialize within their team, a key difference is NEW's emphasis on inter-team collaboration and shared responsibility. Collaboration is potentially present in other specialization or departmentalization schemes, but likely not a centerpiece like it is with NEW.

Despite the growing prevalence of strategic staffing models, little is directly known about their effects on student and teacher outcomes outside of the OC evidence cited above. Previous empirical evidence demonstrates the positive effects of teacher collaboration (Dee et al., 2006; Ronfeldt et al., 2015), and descriptive work on NEW shows positive correlations between teaming and educator outcomes such as satisfaction, evaluation ratings, retention, and sense of autonomy (Ingersoll et al., 2025; Laski, 2024). Other research explores the effects of individual aspects of teamed teaching—including paraprofessionals (Hemelt et al., 2021), peer coaching (Kraft et al., 2018), and distributed leadership (Leithwood et al., 2020)—but does so outside of teamed contexts. One recent study estimates small positive effects of co-teaching, more generally, on student achievement in Massachusetts (Jones & Winters, 2024). To our knowledge, there is currently no direct quantitative evidence estimating the causal impact of comprehensive, team-based staffing models on student achievement.

2.1 The Hypothesized Effects of the Next Education Workforce (NEW) Model

NEW has purported benefits for both teachers and students. For teachers, NEW is intended to reduce the isolation and individual burden of the self-contained classroom by distributing responsibility across a team, theoretically deepening collaboration, allowing for specialization, and expanding professional decision-making authority (Basile et al., 2023; Next Education Workforce, n.d.). These changes are hypothesized to make teachers themselves better off, leading to greater satisfaction and retention. For students, NEW is intended to enable more personalized instruction by allowing teams to differentiate roles and regroup students by need. Teacher-level improvements in working conditions, collaboration, and retention are also hypothesized to translate into gains in student engagement and achievement (Figure 1). Consistent with the teacher-facing half of this logic, descriptive evidence finds that NEW teachers report greater authority and have significantly lower turnover rates than their non-teamed peers (Ingersoll et al., 2025; Laski, 2024). Whether these benefits extend to student achievement is the question our analysis takes up. We use the hypothesized effects described here to structure our analysis.

First, we assess differences by school level. Because organizational structures of schools differ, we hypothesize that strategic staffing models like NEW may operate through distinct mechanisms depending on the school level. Traditional elementary schools often have self-contained classrooms, with one teacher teaching all four core academic subjects to the same roster of students. In these contexts, teaming introduces differentiated roles and specialization to a historically generalized instructional framework. Conversely, middle schools are often departmentalized, and the transition from elementary to middle school is often associated with declines in student achievement and engagement as well as the strength of student-teacher relationships (Anderman, 2003; Marks, 2000; Rockoff & Lockwood, 2010). In middle schools, the team-based model is unlikely to introduce greater specialization than already exists, but it may foster greater collaboration and shared responsibility among already-specialized teachers. Given the fundamentally different baseline structures in elementary and middle schools, we include analyses that separately evaluate impacts of teaming in these different school contexts. Specifically, because middle schools avoid the potential downsides of specialization commonly seen in elementary schools, we might expect positive achievement effects of NEW to be concentrated in middle schools. We also estimate effects separately by test subject given the established pattern in academic literature showing interventions tend to produce larger effects in math than in ELA (Dietrichson et al., 2020).

Second, we examine several channels through which teaming could plausibly raise achievement. In particular, we investigate heterogeneity across student characteristics to assess differential benefits. If NEW's impact operates through increases in differentiation or specialization, we might expect larger impacts on classrooms with more variation in student background, as NEW may make targeting instruction to students that need specialized support more accessible (Jones & Winters, 2024). We also assess differences in team composition and dosage, and assess the possibility of student effects operating through improvements in student attendance, teacher retention, or observed instructional practice. Differences across all of these contexts can provide suggestive evidence of the mechanisms by which teamed teaching can alter student outcomes.

3. Context and Data

We focus our work on one large, diverse school district in Arizona, one of the first districts to implement NEW at scale. The district includes just over 80 schools and approximately 60,000 students and 3,000 teachers in grades Pre K-12. Our panel includes eight years of data on student achievement (2017 to 2025; 2020 is missing because of COVID-19 test cancellations) as well as information linking students to schools, grades, teachers, and courses. All of our analyses limit to core subjects (Math, ELA/Reading, Science, and Social Sciences and History) in grades 3-8. Achievement analyses include only math and ELA.[3] Note that in our district, elementary schools include 6th grade; middle schools include just 7th and 8th grade.

[3] A small portion of our sample—about 3.6% of student-year-course observations—are in multi-grade courses. For multi-grade courses, we assign the modal grade to the course to allow us to create grade-level fixed effects and teaming definitions. Our estimates remain largely unchanged when dropping these courses. We also remove

We identify teamed teachers using data collected by ASU and district staff between 2021 and 2025. Importantly, there is some ambiguity as to what constitutes a teamed classroom, as well as inconsistencies in how the teaming data were gathered over time. In the summers of 2020 and 2021, ASU hosted workshops for teachers planning to implement NEW during the upcoming school year. We use sign-in sheets from those events to proxy actual teacher participation during 2021 and 2022. In 2023 and 2024, ASU gathered the names of teamed teachers from schools via an educator survey. These names were linked to staff IDs by the district. In 2025, teamed teachers were identified by school leadership within the district's administrative data system, and this list was reviewed by ASU staff.

In what we report below, we focus on the most inclusive teacher lists available—essentially treating all teachers that were ever identified as teamed as being in our treated group. NEW and our partner district also provided pared-down lists of teachers or courses that they determined were meaningfully implementing the model to a "moderate" or "definitive" level for every year from 2021 to 2024.[4] Teacher-level teaming rates across definitions and over time are presented in Figure 2, which shows large differences across these three teaming definitions. For instance, in most years, only about half of teachers we count as teamed using the inclusive definition are teamed teachers based on the "moderate" or "definitive" teaming standard. Our models using the most inclusive definition are likely over-identifying teaming rates; if we

observations of students that are in more than one school or grade in a given year or are enrolled in an out-of-grade course; these again represent a very small percent of the total student-year-course observations (less than 5%).

[4] As defined by NEW administration, "moderate" teaming includes "at least two professional educators sharing a common roster of students during the same period of time and making attempts at distributing their expertise, adjusting schedules and dynamic regrouping of students, etc (the team-level elements)." "Definitive" includes "places that were really leaning into Next Education Workforce teaming…not only did they meet the moderate definition of teaming, but we felt like they were effectively distributing expertise, adjusting schedules, and dynamically regrouping students; many of these schools/teams were also receiving direct support from the Next Education Workforce team."

hypothesize that teaming has an effect on student outcomes, this over-identification of treatment will bias our estimates towards zero.[5]

For our student-level analyses, we link teachers to their students using course enrollment data provided by the district and generate a binary indicator for whether a student in a particular subject, grade, and year was assigned to a teamed teacher (always using the most inclusive definition). Our achievement sample includes 2,570 school-grade-subject-year units, and 94% of those units are either fully teamed (12%) or fully non-teamed (82%). A small number (6%) are partially teamed. Partial teaming can happen, for instance, if half of a school's fifth grade math teachers work together on NEW teams while the other half do not. In our preferred specifications, we treat all students in a school-grade-subject-year cell where there is at least one teacher identified as participating in NEW as receiving teaming, even if their own assigned teacher is not identified as teaming. Results therefore capture the effect of NEW implementation within a grade, even if only some teachers within a grade are teamed. Treating teaming as a grade-level variable accounts for the potential concern of student sorting to teamed or non-teamed teachers within a grade. For example, if a principal systematically sorted their students with the most growth potential within a grade to teamed teachers' classes, comparing these students to non-teamed students in the same grade would incorrectly assign that growth as an effect of teaming. If we hypothesize a positive effect of being assigned to a teamed teacher, the models identifying teaming at the school-grade-year-level would then generate conservative lower-bound estimates of the NEW effect by including untreated students in the treated group. These models also account for the possibility that NEW has positive or negative spillover on technically untreated students in the same grade. Our findings are qualitatively similar when

[5] Our pattern of results on student outcomes is broadly similar when using the most restrictive teaming definition, but estimates are slightly closer to zero and less precisely estimated.

using an alternate identifier based only on whether a student is directly assigned to a teamed teacher, given that most school-grades with at least one teamed teacher are fully teamed.

3.1 What Teaming Looks Like

Between 2021 and 2025, NEW was implemented across 52 different schools (about two-thirds of the district), in grades K-12, and across all core subjects. This section describes NEW's implementation within math and ELA in grades 3-8. Table 1 describes that implementation using data aggregated to the school-subject level, by school level and year; Panel A reports the proportion of schools with at least one teamed teacher in a given school level and subject and Panel B the proportion of students in teamed units who are assigned directly to a teamed teacher. Importantly, these panels highlight the staggered rollout of NEW across schools, grades, and subjects over time.

Across school levels and subjects, NEW was implemented in about 12% of schools in 2021 (Panel A). Implementation was uneven over the next two years but roughly doubled in 2024 relative to 2023. Elementary and middle school rates diverged in 2025: elementary implementation declined to roughly its 2021 level, while middle school implementation remained much closer to its 2024 level. In the final year of our sample period, middle schools were more likely to implement than elementary schools.[6] Panel B presents the proportion of students assigned directly to teamed teachers within grades that are teamed. In any teamed school-grade-subject-year, between 65 and 100 percent of students are assigned directly to a teamed teacher, though rates vary by year.

[6] As noted in section 3, teaming data was collected in a notably different way in 2025: previous years used teacher sign-ins or survey responses, while in 2025, teamed teachers were identified by school leadership within the district's administrative data system using criteria provided by NEW. This change coincides with a decline in the number of teamed teachers (Figure 2), which could represent a real change in teaming rates or be a byproduct of the different method of data collection. Our preferred specifications treat a grade as teamed every year after initial implementation, so this is not a major concern for identification.

Table 2 describes team composition by school level using data from 2023 and 2024. These data are limited in two respects: they cover only these two years, and they identify who is on each team rather than which of a teacher's courses (and students) are actually teamed. We use the data to describe the team members themselves. Values represent all teamed teachers for the average school in a given school level and year. On average, schools had about four teams, and teams had about four members across two unique roles. In 2024, despite an increase in the total number of teams across schools, teams became smaller and less diverse in terms of unique educator roles. Specifically, average team size fell from roughly 4.1 members spanning 2.4 distinct roles in 2023 to 3.7 members spanning 1.8 roles in 2024. Teams were also less likely to feature lead teachers in 2024 than in 2023. Teacher candidates were relatively uncommon but shifted towards middle school in 2024, possibly reflecting temporal variation in schools' capacity to host pre-service teachers rather than any practices specific to teaming. Elementary teams were the only teams with general paraeducators, while half of middle school teams featured a special education teacher. Otherwise, elementary and middle school teams seemed generally similar based on the limited team composition data we have available.

Our data do not identify which specific components of NEW are practiced by the teachers on each team. "Teaming" represents a spectrum of potential collaborative activities, from non-instructional tasks such as common planning time, to instructional tasks including interdisciplinary lessons and collaborative teaching. Importantly, schools adapt NEW and its individual aspects to their own contexts, given feasibility. NEW most likely formalizes and increases collaboration between teachers and expands opportunities for differentiation.

NEW models were first implemented in some grades in the district in 2021, the first full school year of the COVID-19 pandemic. All schools in our study district began that year in virtual learning. Between September and December, all schools operated hybrid learning models, featuring a blend of in-person and remote learning. Overall, all schools spent at least 80% of the school year (months) in hybrid learning (COVID-19 School Data Hub, 2022). While notable, this shift in instructional context does not threaten our identification strategy, because school instructional mode did not vary systematically with teaming status: NEW and non-NEW grades followed the same hybrid schedule, and the mode was set at the school level rather than by grade or classroom.[7]

Table 3 presents summary statistics comparing students and teachers that implemented NEW to those that did not. For consistency, the unit of analysis for this table is the school-grade-year, but results are similar when considering different units. We note several meaningful differences between grades that did and did not implement NEW. Implementing grades have demographically different students; students in NEW grades are significantly more likely to identify as Hispanic or English Language Learner and significantly less likely to identify as White. Students in NEW grades also have significantly lower achievement scores and are more likely to be identified as chronically absent (defined as having an attendance rate lower than 90%). Finally, we also note that teachers working in NEW models have significantly less experience than non-teamed teachers. Broadly, NEW is implemented disproportionately in lower-performing grades with less-experienced teachers, on average.

4. Method

Our study design capitalizes on the staggered implementation of NEW across schools and grades beginning in 2021. This variation enables a difference-in-differences (DiD) design that

[7] Because teaming is not associated with instructional mode, our identification strategy remains sound through COVID-19 instructional changes. That said, virtual instruction could change the treatment itself: teaming likely works differently in a virtual environment. However, in supplementary analyses where we drop 2020-21, our findings do not change qualitatively, suggesting that instructional mode is not a concern to interpretation.

compares changes in student performance over time between school-grade cells that have implemented NEW and those that have not yet implemented NEW in a given year (or that have discontinued implementation).

A causal interpretation of the DiD estimates relies on the parallel trends assumption: absent NEW, outcomes in treated school-grade cells would have followed the same trajectory as those in comparison school-grade cells, despite non-random selection into treatment. Put differently, the decision to implement NEW in a particular grade was not driven by that grade's expected future performance. To assess the plausibility of this assumption, we present tests of pre-treatment trends below.

We define treatment at the grade-subject level: a student-grade-subject-year cell is considered treated in year *t* if at least one of the teachers in that cell is participating in NEW in year *t*. We start with a baseline canonical two-way fixed effects (TWFE) regression of the following form:

$$A_{isgt} = \beta Teamed_{sgt} + \gamma_{sg} + \delta_t + \varepsilon_{isgt} \quad (1)$$

for student *i* in school s, grade g, and year *t*. A_{isgt} represents the standardized test score from grades 3-8 for student *i* in school *s* in grade *g* in year *t.* Test scores are standardized within year-grade-subject cells using the district-wide sample, and standard errors are clustered at the school-by-grade level. $Teamed_{sgt}$ is a binary equal to 1 for all students in a school-grade-year cell *sgt* where there is at least one teacher identified as participating in NEW during a given year. As noted above, defining treatment at the school-grade-year level accounts for the possibility of within-grade sorting between teamed and non-teamed classrooms. For example, if a principal systematically assigned students with the greatest growth potential to teamed classrooms, comparing those students directly with non-teamed students in the same grade would incorrectly attribute their higher growth to NEW rather than to selection.

If NEW has a positive effect on students assigned to teamed classrooms, our school-grade-year specification yields conservative, lower-bound estimates of its impact because the treated group includes some students who were not directly exposed to teaming. Consequently, the estimated effects capture the net impact of NEW implementation within a grade, including any positive or negative spillover effects on students in non-teamed classrooms. Our findings are qualitatively similar when treatment is instead defined at the student level, where $Teamed_{it}$ is equal to 1 if student i is directly assigned to a teamed teacher in year t.

Recent literature on difference-in-differences estimates raises concerns about bias in two-way fixed effects estimates (Goodman-Bacon, 2021; Callaway & Sant'Anna, 2021; Roth et al., 2023). The canonical TWFE model could be biased by at least three unique aspects of our study context: the staggered rollout of NEW over time, the possibility of heterogeneous treatment effects over time, and the fact that grades are not consistently treated or untreated (and can move from treated to untreated over time). In these situations, the TWFE estimator can place negative weights on some cohort-time treatment effects, producing estimates that do not have a clear causal interpretation as a weighted average of treatment effects (Goodman-Bacon, 2021; de Chaisemartin & D'Haultfœuille, 2020). This arises because TWFE implicitly uses already-treated units as comparisons for newly-treated units, the so-called "forbidden comparisons" problem (Roth et al., 2023).

Our preferred specification addresses these concerns by using the imputation-based estimator of Borusyak, Jaravel, and Spiess (2024; henceforth BJS). The BJS estimator proceeds in two steps. First, the procedure estimates the school-by-grade and year fixed effects γ_{sg} and δ_t

using only observations of untreated school-grades (both never-treated and not-yet-treated). Second, for each treated observation, the estimator imputes a counterfactual outcome under no treatment based on the estimated fixed effects, and the treatment effect is computed as the difference between the observed outcome and the imputed counterfactual. The average treatment effect on the treated is then computed as a weighted average of these unit-level treatment effects. Because the fixed effects are estimated only from untreated observations, the estimator avoids forbidden comparisons by construction.

The BJS estimator is ideal for our context because it naturally accommodates heterogeneous treatment effects across cohorts and time since initial implementation, allowing us to examine dynamic treatment effects without imposing functional form assumptions. In addition, the estimator produces formal placebo tests for parallel trends in the form of estimated pre-period coefficients, which can be used both visually and via joint hypothesis tests to assess the validity of the design. Following the recommendation of Sant'Anna and Zhao (2020), we estimate our preferred specification without any control variables. Including time-varying covariates requires assuming that the covariates are themselves unaffected by treatment, an assumption we cannot defend in our context, given evidence, presented below, that student composition shifts systematically with NEW adoption.

To assess robustness to alternate specifications, we also estimate our main results using the method developed by de Chaisemartin and D'Haultfœuille (2020; henceforth dCDH), which uses a first-differencing approach to recover heterogeneity-robust average treatment effects. This estimator makes weaker assumptions than BJS about the structure of treatment effect heterogeneity over time, accommodating school-grades switching into and out of treatment, which is common in our setting: 70% of treated school-grades change from treated to untreated at some point, and the majority of these changes are driven by the shift in treatment definition between 2024 and 2025. Because it is unclear whether this shift in definition represents a true treatment exit (see footnote 7), we prefer to use an absorbing definition of treatment like the BJS estimator (in other words, once a unit is treated, we consider it treated until the end of the panel). The dCDH estimator nonetheless serves as a useful robustness check against alternative weighting schemes across the cohort-time treatment effect distribution.

In the appendix, we additionally report results from three further estimators: the Gardner (2022) two-stage imputation estimator, which is closely related to BJS in spirit but uses a different two-stage procedure; a stacked difference-in-differences design following Cengiz et al. (2019); and the Callaway and Sant'Anna (2021) estimator both with and without grade-size weights.

Selection into NEW is a primary identification concern in this context. Because participation is voluntary, schools or principals opting into the treatment may possess unobserved characteristics, such as higher motivation or greater administrative capacity, that are also correlated with student achievement. Our preferred specifications include some level of school fixed effects to account for this concern, essentially limiting to comparisons within schools, but implementation could still potentially coincide with principal turnover within the school. If so, treatment estimates could reflect the impacts of new leadership rather than NEW. Our data do not support this interpretation: among implementing schools, only 5.4% of initial adoption years coincide with the arrival of a new principal (compared to roughly 20% in non-adoption years). NEW adoption is associated with leadership stability rather than turnover, so incoming leadership is unlikely to explain any estimated effects.[8]

[8] Pre-adoption turnover is similar, ranging from 12 to 24 percent across the six years prior to adoption.

Teachers within implementing schools may also opt into NEW, or principals may choose to implement NEW in grades where teachers are more willing or motivated to try something new. We address this potential for school- and grade-level selection by incorporating school-by-grade fixed effects (γ_{sg}). These account for persistent, unobserved heterogeneity at the school-grade level, such as the systematic assignment of high-performing teachers to specific grades or stable differences in grade-level resources. Consequently, our identification relies on within-school-grade variation in the timing of NEW adoption. We combine these with year fixed effects (δ_t), which absorb idiosyncratic, contemporaneous shocks common to all schools and grades in a given year, such as the disruptive effects of the COVID-19 pandemic or state-level policy shifts. Our preferred specification restricts the comparison to changes occurring within a specific school-grade over time, relative to the trends of non-treated school-grades over the same time period. This approach ensures that our estimates are not driven by time-invariant school-grade characteristics or year-specific volatility.[9]

4.1 Assessing Potential Sources of Bias

We test for differential trends in the composition of treatment and comparison school-grades by running a modified event study model, where each characteristic is included as the dependent variable and pre-period years are captured by a continuous linear variable rather than a set of indicators. The coefficient on this variable can be interpreted as testing whether treated school-grades had characteristics that were on different trends than those of untreated school-

[9] A natural alternative would replace year fixed effects with school-by-year fixed effects, absorbing all time-varying shocks common to a school. We do not adopt this specification because it meaningfully changes the interpretation of our results. In models with school-by-year fixed effects, NEW's impact can only be identified from within-school-year contrasts in teaming across grades. Because many schools—particularly middle schools, which span only 7th and 8th grade in our district—adopt teaming across all tested grades in the same year, these schools contribute no within-school, within-year cross-grade variation and drop out of identification entirely. The resulting estimate would be recovered from a non-representative subset of partially adopting schools. Specifically, 19% of school-years with any teaming are teaming school-wide, and thus would not contribute to the impact estimate in models with school-by-year fixed effects.

grades in the years prior to initial treatment. Table 4 presents this information for each school level and subject, with each cell in the table representing output from a separate regression. Insignificant coefficients indicate that treated and untreated units were not on differential trends in the years prior to implementation for a given characteristic.

We find that five of our ten tests are significant in elementary math (column 1) suggesting significant violation in pre-treatment trends. This is less so at the middle school level in math, as only student enrollment is significant. Student enrollment is also significant in elementary ELA. In middle school ELA, while enrollment levels are not significant, we find significant differences in English Language Learner status and the average daily attendance of students. Overall, the evidence of compositional shifts before implementation raises concerns about potential bias in our estimates, particularly in elementary math, which we interrogate further in the subsequent sections.

We also test whether the characteristics of treated units in our baseline year (2017) predict initial adoption years. Estimates are presented in Table 5. We find few significant differences on individual characteristics, though estimates differ across grades and subjects. The clearest pattern is in baseline achievement: elementary math and middle school ELA grades with higher achievement at baseline adopted NEW significantly later, indicating that early adopters had lower baseline achievement than later cohorts. We see no analogous relationship in middle school math or elementary ELA, where the baseline achievement coefficients are positive but statistically insignificant.

To assess whether baseline characteristics jointly predict adoption timing, we report two tests for each cell: one over the full set of covariates and one over a parsimonious set of six characteristics relevant to selection. We caution against over-interpreting these joint tests in middle school, where only 12 (math) and 17 (ELA) treated grades leave few residual degrees of freedom to reliably test these patterns. In the elementary cells, which are more adequately powered, the joint tests are sensitive to how demographic composition is included, as evidenced by notable changes in p-values across the two different F tests.

We read these results as corroborating, rather than driving, our identification strategy. The small number of treated middle school grades makes cohort-balance tests inherently underpowered, so we place greater weight on BJS pre-trend tests in the following section as evidence on the credibility of the middle school estimates. To the extent baseline achievement predicts adoption timing in other cells, these results underscore the importance of the BJS estimator, which accounts for differences in treatment effects across early and late adopters.

4.2 Assessing Parallel Trends

The identifying assumption underlying our preferred specification is that, absent NEW, outcomes in treated school-grades would have evolved in parallel to outcomes in untreated school-grades. We assess this assumption in two ways: visually through event studies that plot pre-period coefficients estimated in both the canonical TWFE model and the BJS procedure, and formally through joint hypothesis tests of pre-period parallel trends of the BJS estimator. P-values of these joint tests are included in Table 6. As discussed above, we have theoretical reasons to expect heterogeneous treatment effects across elementary and middle schools and subjects, given findings in the previous literature. Thus, we conduct these tests separately for elementary (grades 3-6) and middle school (grades 7-8) grades and for math and ELA achievement.

Figures 3 and 4 present event studies from the canonical TWFE specification for achievement outcomes. Figure 3 pools both subjects across all grades, while Figure 4 disaggregates by school level. Two patterns emerge. First, pre-trends vary across school levels and subjects, consistent with the heterogeneity documented above. Second, the pre-period coefficients are not uniformly supportive of parallel trends: while middle school math displays clean pre-trends, the event studies for ELA overall, math overall, middle school ELA, and elementary math each exhibit at least suggestive downward pre-treatment trajectories.

We interpret these patterns as grounds for caution rather than as conclusive evidence against the design's validity. As noted above, the canonical TWFE event study is itself susceptible to bias under staggered adoption with heterogeneous treatment effects, meaning that apparent pre-trends in the TWFE specification may reflect "forbidden comparisons" rather than genuine violations of parallel trends. We therefore treat these TWFE event studies as a conservative diagnostic tool and rely on the BJS imputation estimator—which is robust to treatment effect heterogeneity under staggered adoption—for both our primary estimates and the formal pre-trend tests reported below.

Figures 5 and 6 present BJS event studies for math and ELA. Figure 5 displays both subjects across all grades in our sample, while Figure 6 displays both school levels and subjects. In math, while individual point estimates are not statistically distinguishable from zero, both the overall and elementary figures show some suggestive evidence of downward pre-trends, which could bias our estimates. The joint test of pre-period parallel trends is not significant for math overall ($p = .108$), but is significant for elementary math ($p = .036$).[10] We thus emphasize caution in over-interpreting the math elementary results given this significant pre-trend, as we cannot confidently rule out that any apparent treatment effect in elementary grades reflects continuation

[10] We use the pre-trends test native to the did_imputation command in Stata, which requires a user-determined input of the number of pre-periods to include in the test. We present the p-values from the test for parallel pre-trends using five pre-periods, but also present the p-values for every option in Appendix Table 1.

of this pre-existing trend. The middle school math pre-trend is markedly different, without a consistent downward pre-trend. All individual coefficients are indistinguishable from zero, and the joint test of coefficients confirms we cannot reject parallel pre-trends ($p = .836$). For middle school math, we have more convincing evidence of parallel trends, both visually and via formal testing. Reassuringly, we also see no sign of an Ashenfelter-type dip: the coefficient immediately before implementation is in line with the other pre-period estimates.

In ELA, both the overall and middle school event studies show suggestive evidence of pre-trends, though magnitudes are small. Both joint tests reject parallel trends ($p = .035$ for ELA overall and $p = .006$ for ELA middle). Again, we emphasize caution in interpreting results from these subsamples given the fact that pre-trends are not parallel. Elementary ELA does not have a clear pre-trend, and the joint test cannot reject that they are parallel ($p = .773$).

However, these tests may be underpowered, given imprecision in our event study coefficients. In particular, it is possible that the suggestive post-period effects are explained by linear trends continuing from the pre-period. We test this possibility using the diagnostic proposed by Roth (2022). We first calculate the smallest linear pre-period trend that we can detect with 80% power. We then overlay this trend onto our event study coefficients, which we plot in red in Figure 7. Results show that a hypothetical pre-period trend threatens our interpretation in elementary ELA, even though previously described tests cannot reject parallel trends: the post-period trends cannot be differentiated from the possible pre-period trend in red. Estimates for other grades and subjects are not similarly threatened, as post-period trends clearly diverge from the pre-period. In middle school math, because the estimated impact is positive, we confirm that the pre-trend is too flat to adequately explain the post-period upward trend. Appendix Table 2 includes additional diagnostic statistics from this analysis.

The differential pre-trend evidence across cells has direct implications for how we interpret our results. We are most confident in the parallel trends assumption for middle school math, where both visual and formal tests support the design. We are less confident in the assumption for the remaining grades and subjects. In those cells, bias from pre-existing trends is likely, and the direction of bias is unpredictable: pre-trends could either inflate or deflate the apparent treatment effect depending on whether the trend continued or reversed after NEW adoption. In short, all pre-trend tests are satisfied only in our middle school math cell; we present other cells with explicit caveats about the limitations of identification.

5. Primary Findings

The primary achievement results are presented in Table 6. Each cell in each panel represents output from a separate regression model. All models include fixed effects for school-by-grade and year. In columns 1-3, the outcome variable is math achievement. In columns 4-6, the outcome variable is ELA achievement. Achievement is standardized within year-grade-subject cells using the district-wide sample. The first and fourth columns include our full sample of grades 3-8. The second and fifth columns limit to elementary schools (grades 3-6); the third and sixth columns limit to middle schools (grades 7-8). Each model estimates how differences in teaming over time within a given school-grade are associated with achievement.

In both math and ELA, the overall impact estimates (columns 1 and 4) are near-zero, insignificant, and inconsistently signed across estimators. In short, we find little evidence of an overall impact of NEW on achievement in either subject.

The effects of NEW are, however, heterogeneous across grades and subjects. Middle school math—where the pre-trend evidence is most convincing and a plausibly causal interpretation is best justified—also shows the most consistent impact estimates across specifications: all estimates are positive, of roughly similar magnitude (ranging from 0.06 to 0.10

SD), and two of the three are statistically significant at conventional levels ($p < .05$) or marginally so ($p < .10$).

This pattern holds in Appendix Table 3, which presents results from several alternative estimators. For middle school math, all estimates remain positive, and the Gardner (2022) and stacked estimates are of comparable magnitude to our preferred BJS estimates. The Callaway and Sant'Anna estimates are somewhat smaller and less precisely estimated, which we attribute to those estimates being constructed from collapsed school-grade-year level data rather than student-year level data.

The above findings suggest NEW had a moderately large and meaningful impact on middle school math achievement, on the order of 0.08 SD. Assuming a one standard deviation change in teacher effectiveness represents roughly 0.15 student standard deviations in achievement (Hanushek & Rivkin, 2012), this impact is equivalent to roughly 53% of a teacher standard deviation. Put differently, the estimated impact is equivalent to moving from a median teacher to a teacher at roughly the 70th percentile of the teacher effectiveness distribution, or moving from a novice teacher to a teacher entering their second year of teaching in our data. The estimates are credible given the pre-trend tests presented above and consistency across estimators. The concentration of effects in middle school math also aligns with our theoretical framing on NEW's impacts differing across school levels (see Section 2) and with a broader pattern in which academic interventions tend to produce larger effects in math than in ELA (Dietrichson et al., 2020).

Estimates in other grades and subjects are less convincing causally, but we present them for transparency. In elementary math, we find some evidence of significant and borderline significant negative effects, which we interpret with caution. The elementary math estimates are not consistent across estimators—the dCDH estimate is actually slightly positive—and the BJS joint test of pre-period parallel trends is significant (p = .036), so we cannot rule out that the apparent effect reflects a continuation of a pre-existing downward trend. In elementary ELA, by contrast, the BJS estimate is significant and negative, and we cannot reject parallel pre-trends (p = .773). We nonetheless treat this cell cautiously as well, as the estimates are again directionally inconsistent across estimators.

The negative elementary point estimates are directionally consistent with Fryer's (2018) finding that introducing specialization into historically self-contained elementary classrooms has a negative impact on achievement. Estimates in the remaining ELA cells (overall and middle school) are inconsistent across specifications and lack strong evidence for parallel pre-trends, so we again refrain from causal conclusions.

5.1 Alternate Specifications

A potential concern with our preferred specification is that teachers with different unobserved attributes are correlated with selection into NEW, either by opting in themselves or by their principal's decision to place them on a team. To account for this potential concern, we also present alternate specifications utilizing teacher fixed effects, as follows:

$$A_{isgt} = \beta Teamed_{sgt} + \theta_j + \delta_t + \mu_s + \varsigma_g + \varepsilon_{isgt} \quad (2)$$

where θ_j represents teacher fixed effects. These models allow us to estimate differential impacts on student outcomes within-teacher, essentially comparing a teacher's performance before and after working on a NEW team. These specifications address concerns about our impact estimate potentially being contaminated by unobserved characteristics of NEW teachers.

Table 7 presents results from these specifications, following equation (2). Broadly, estimates are quite similar to our preferred specifications. Estimates across achievement subsamples are generally near-zero and non-significant, with one exception: in middle school math, we again see evidence of a positive impact of NEW. The estimated within-teacher impact of teaming is 0.067 SD, and this estimate is marginally significant ($p = .062$). Taken together, the fact that these patterns align across models is reassuring that our estimates are not substantially biased by selection on unobserved characteristics at the teacher level.

Another potential concern is selection at the *student* level: teachers or parents could, for instance, sort students into or out of NEW classrooms based on unobserved factors that influence their achievement growth. Appendix A presents alternate specifications using student fixed effects. Estimates are uniformly small and statistically insignificant. The middle school math estimate is nearly zero (0.004 SD), but the confidence interval is wide enough to contain both zero and estimates from our preferred estimators (roughly 0.08 SD). We interpret these findings with caution given concerns outlined in the appendix.

Finally, we also test whether our preferred estimates are robust to alternate sample restrictions. We focus on two specific data decisions that may warrant scrutiny: the structural change in how treatment status was recorded in the 2025 academic year and the ability of treated units to stop treatment. Appendix B presents alternate models for the achievement analysis, testing the robustness of our estimates across more restricted samples in our preferred BJS estimator. Findings remain qualitatively similar, suggesting that our preferred estimates are not disproportionately driven by decisions to include observations from 2025 and grades that stop teaming.

6. Robustness and Heterogeneity Checks

6.1 Additional Threats to Interpretation: Teacher Assignment and School Leadership

While we have interpreted our results thus far to represent the impact of teaming itself, a natural concern is that any estimated effects could actually be the result of teacher reassignment to teaming grades. Specifically, if principals form teams by reassigning their most effective teachers into treated grades, any effects we attribute to NEW could reflect that compositional shift rather than an impact of the teaming model itself. We examine this possibility descriptively by calculating how teacher assignments change as grades begin teaming. Appendix Table 4 reports the share of teachers whose grade-and-subject assignment matches the prior year ("stayers," column 1) and the share who are new to that assignment ("newcomers," column 2) for each school level, subject, and teaming combination. Newcomers can also be decomposed by whether they came from within the same school (column 3), another school in the district (column 4), or were not in the district in the previous year (columns 5-7). The latter group includes multiple types of teachers: those that had previous experience in the district but were not working in the previous year, likely due to taking leave (column 5), first-year teachers (column 6), and incoming experienced teachers that did not previously work in the district (column 7).

Two patterns emerge. First, the onset of teaming generally coincides with modestly higher rates of newcomers to a grade and subject. Relative to both never-teamed grades and pre-teamed grades, grades in their first year of teaming generally have a smaller proportion of stayers, which would be consistent with principals reorganizing staff just as they kick off teaming. Second, the source of teaching churn differs across grades and subjects. In particular, in middle school math, the rise in newcomers is driven almost entirely by teachers new to the district's administrative data (columns 5-7), and within-school reassignment is flat (column 3). In other grades and subjects, the increase comes mainly from within-school reassignment.

While descriptive, this evidence suggests that within-school teacher sorting is likely not driving the middle school math result. Specifically, the majority of middle school math newcomers were not teaching in the school or district in the previous year, suggesting principals likely were not selecting them to team based on demonstrated effectiveness. This is particularly interesting given strong evidence that the arrival of new teachers typically decreases student achievement (Atteberry et al., 2017). The fact that middle school math achievement rose with teaming alongside an influx of teachers that were new to the school would be consistent with teaming functioning as a support structure for newcomers rather than a vehicle for concentrating already-effective teachers. That said, we stress that this analysis is descriptive and relies on a small number of observations, so we offer it as a hypothesis for future research.

A related concern is that the estimated effect reflects the arrival of a new principal who both adopts NEW and improves achievement, rather than teaming itself. We find this unlikely. As reported in Section 4, NEW adoption coincides with a new principal in only 5.4% of initial adoption years, well below the roughly 20% rate in non-adoption years; if anything, adoption is associated with leadership stability.

6.2 Heterogeneity of Effects Across Students

We next examine whether NEW's effects concentrate among particular student subgroups. We treat this as an exploratory description of where effects appear rather than direct evidence on mechanism; a more targeted test of instructional differentiation appears in Section 7.

To probe heterogeneity in impacts, we focus on our preferred BJS estimator and estimate effects across various student characteristics and the novice composition of a grade's teaching staff in 2021.[11] Because we identify treatment at the school-grade level, these analyses

[11] We estimate effects using the hetby() option within the did_imputation command in Stata, which recovers a separate average treatment effect for each subgroup while preserving a joint variance-covariance matrix, allowing us to test the difference between subgroups directly. So that each subgroup is compared against its own counterfactual rather than the grade-wide average, we interact each subgroup identifier with both the school-grade and year fixed effect. Without this interaction, a subgroup's estimated effect mechanically absorbs its baseline achievement gap relative to the rest of the grade. We fix each student moderator at its pre-treatment value to avoid conditioning on characteristics that NEW may itself affect, and we measure novice composition in 2020-21.

essentially assess whether the effect of being in a teamed grade differs across different types of students within that grade.

We treat these results as exploratory. Our power to detect differences is limited, as splitting grade-level treatment into student subgroups divides our estimation sample. A true subgroup difference of plausible magnitude would be unlikely to clear conventional thresholds at this sample size, so we read insignificant differences as undetectable rather than absent. In addition, our most credibly identified cell, middle school math, is our smallest. In these smaller cells, the leave-out variance correction we use elsewhere is not computable, so we instead report standard errors from a more conservative pooled-variance specification.

Table 8 presents results from these analyses. Across most cells, the estimated differences between subgroups are small and statistically indistinguishable from zero. We take this as suggestive evidence that NEW's effects are generally shared across student populations and not driven by gains from any single group. Focusing specifically on middle school math, our most credibly identified subsample, we find that estimates across all five indicators are positive and of comparable magnitude. In other words, the middle school math gain does not seem to be driven by any single group. We do note that Hispanic students have somewhat larger estimated impacts (0.116 SD compared to 0.074 SD for non-Hispanic students), but the difference is not statistically significant ($p = .165$). We also find suggestive evidence that impacts are much larger in grades with novice teachers, though the difference is again not significant at traditional levels (Difference = 0.125, $p = .143$).

We also note that, in both overall and elementary math, ELL students realize significantly larger impacts than non-ELL students. Given the weaker identification in these subsamples, we interpret these differences cautiously. Overall, effects appear broadly shared across student populations, with some imprecise indication that more historically underserved groups may benefit slightly more. These small differences are consistent with, though not strong enough to confirm, a targeted-support mechanism, where historically underserved students might disproportionately benefit from NEW because they could gain the most from differentiated support.

7. Mechanisms

The analyses above identify a positive effect of NEW concentrated in middle school math. Because our data do not capture the in-classroom experiences of students directly, the mechanisms underlying this effect remain uncertain. In this section we examine several channels through which teaming could plausibly raise achievement and which may operate differently across elementary and middle schools: instructional differentiation, team composition, teaming dosage, student attendance, and the effectiveness and stability of the teaching workforce. As we detail below, several of these channels can be ruled out, but the available evidence does not permit us to isolate a clear mechanism—a question we leave to future work.

7.1 Instructional Differentiation

The most direct channel through which teaming could raise achievement is instructional differentiation. By distributing responsibility across a team, NEW may lower the marginal cost of tailoring instruction to student need; theoretically, teams could group and re-group students across educators and target support more easily than a single teacher. If differentiation is key to improved student outcomes, we would expect larger effects where the scope for it is greatest, such as in classrooms serving students with more heterogeneous incoming achievement. We assess this possibility by incorporating the lagged achievement of incoming students in teamed grades. We first look descriptively and find that incoming achievement dispersion is broadly similar across grade-subject cells: the average within-unit standard deviation of prior-year scores ranges narrowly from about 0.73 to 0.83 across the six subgroups, and middle school math, where the achievement effect is most consistent, actually has the lowest level of dispersion (0.73). Thus, differences in the incoming student heterogeneity alone cannot explain why effects concentrate in middle school math.

To probe the differentiation channel more directly, we use each school-grade-subject's incoming dispersion—defined as the standard deviation of students' prior-year scores in its first teamed year—and interact this measure with teaming status in our preferred middle school math specification. Consistent with a differentiation channel, the estimated teaming effect rises with a school-grade's incoming dispersion: it is near zero for units at the 25th percentile of incoming dispersion (about 0.01 SD), roughly 0.09 SD at the median, and about 0.13 SD at the 75th percentile. The coefficient on the interaction is positive but imprecisely estimated ($p = .111$). Overall, this suggests teaming was more impactful in more heterogeneous middle school classrooms, though we note again this is estimated imprecisely off of a small sample. Future work should further investigate this as a potential mechanism.

7.2 Team Composition

The composition of a team may shape how teaming works and therefore what it does for students. Who is on a team may drive the focus of collaboration: a team whose teachers share a grade may more easily coordinate around a common roster of students, while a team spanning grades might focus more on ensuring cohesive instruction across grades. A team that includes non-teachers likely distributes instructional work differently than one made up only of teachers of record. These are distinct collaborative arrangements, and they could plausibly differ in how effectively they translate teaming into student learning. We explore this by examining a limited dataset on team composition.

For two years of our panel, we observe which team each teacher belonged to, allowing us to explore how team structure differs between elementary and middle schools. Importantly, we do not have data on which *courses* a teacher shared with team members: for example, a middle school math teacher may teach both 7th- and 8th-grade classes but team only with 7th-grade colleagues; in our data we can only identify them as a 7th- and 8th-grade teamed teacher. With this limitation, we can classify a team as "single-grade" when all members predominantly teach the same grade. Using this definition, we find a clear difference across school levels: 74% of elementary teams are single-grade, compared with 48% in middle school. Middle school teams are thus more likely to contain teachers who teach different grades. We caution, however, that this need not mean middle school teams collaborate across grades. Middle school teachers are more likely to teach multiple grades, so the contrast may reflect these teachers' broader teaching loads rather than genuinely cross-grade teaming. That said, the possibility that cross-grade collaboration is more common in middle school is real, and worth exploring further with better data. We also find that elementary teams are somewhat more likely to include non-teacher members, such as paraeducators, though the data on educator roles are less complete than grade-level data.

We cannot draw direct conclusions from these data: team identifiers cover only part of the panel, we observe team membership rather than which courses are teamed, and we lack information on the within-team practices that might distinguish cross-grade from within-grade collaboration. We offer this as descriptive evidence that team composition may matter and should be a focus of future work.

7.3 Team Dosage

A third possibility is that the middle school math effect reflects longer exposure to teaming rather than something specific to that setting. Our event study figures suggest there may be benefits to increased dosage: the coefficients three and four years after implementation are larger than those in the earlier years, though their confidence intervals overlap. If middle school math teams had simply been teaming longer than teams elsewhere, greater accumulated exposure could be driving our result rather than anything specific about the subject-grade context. We assess this descriptively by plotting, for each school level and subject, the distribution of school-grade units across years since initial implementation (Figure 8). Because a school-grade contributes to this distribution in each year it is teamed, the distribution reflects how long units have been teaming across our panel.

The distributions are broadly similar across school levels and subjects: in every cell, roughly half of teamed unit-years are first-year implementations (46-52%). Middle school math looks similar to other grades and subjects; in fact, middle school math actually has the highest share of first-year units (52%) and among the lowest at four years (13%), the opposite of what a longer-exposure explanation would require. We therefore find no indication that our middle school math results are an artifact of those teams having been in place longer; teaming tenure is distributed similarly across the grade-subject cells we compare. Note that this check addresses whether exposure differs across contexts, not whether the effect itself grows with exposure, which we cannot yet resolve given the relative imprecision in our event study specifications.

7.4 Student Attendance

NEW may have an impact simply by increasing student attendance. To assess this potential mechanism, we estimate impacts of NEW on student attendance. For our attendance analyses, we run similar models to our achievement analysis, with the outcome of interest being attendance rate (defined as the proportion of days attended in a given year, ranging from 0 to 1). Our attendance analyses also limit to students that were enrolled at a school for at least 40 days, which includes over 97% of student observations.[12]

Results are in Table 9. Again, each cell in each panel represents output from a separate regression model, and columns represent different grade bands. As seen in Appendix Figures 1 and 2, we cannot make convincing claims about parallel trends in attendance. Estimates are consistently negative but very small—the largest is about one percentage point, or fewer than two days in a 180-day school year—and the teacher-fixed-effects specifications (Table 7, columns 7-9) are similar. We therefore find no evidence that changes in attendance explain the middle school math effect; given the weak identification, we read this as making the attendance channel unlikely rather than conclusively excluding it.

7.5 Teacher Retention and Evaluation Ratings

We next turn to teacher-level outcomes to assess whether changes in teachers could explain the patterns in our achievement results. First, we test how NEW impacts teacher turnover, which has demonstrated negative effects on student achievement (Atteberry et al., 2017; Ronfeldt et al., 2013). Second, we assess NEW's impact on teacher evaluation ratings, an imperfect but informative proxy for teacher effectiveness. Depending on their experience and previous ratings, teachers in our district are observed one to two times annually using the Charlotte Danielson framework. We observe their overall annual scores, which incorporate both classroom observations and measures of student or school performance. Ratings range from ineffective (1), developing (2), effective (3), or highly effective (4). Ratings are highly skewed,

[12] Appendix Figures 1 and 2 present event studies for student attendance. Violations of parallel pre-trends are stark in the attendance sample, and all BJS joint tests of pre-period parallel trends are statistically significant at the 1% level. Thus, while we present attendance results here, we do not find convincing evidence that our estimates accurately capture the causal impact of NEW.

with just 5% of teachers receiving an ineffective or developing rating. Even so, previous research has demonstrated that teacher evaluation ratings are predictive of both teacher performance and student achievement (Bacher-Hicks et al., 2019; Kane & Staiger, 2012).

For teacher-level outcomes, we run models of the following form:

$$A_{jsgt} = \beta Teamed_{jt} + \gamma_{sg} + \delta_t + \varepsilon_{jsgt} \tag{3}$$

for teacher *j* in school *s*, grade *g*, and year *t*. $Teamed_{jt}$ is a binary equal to 1 if a teacher *j* is on our teamed teacher list in year *t*. Standard errors are clustered at the school-by-grade level.

Table 10 presents our results on teacher outcomes, following equation (3). Columns 1-3 present impacts on turnover, which we define as a binary equal to 1 if a teacher is no longer teaching in their school in year *t+1*. We find that, in the most basic model, turnover rates among teamed teachers are 7 to 8 percentage points lower than their non-teamed peers, and this difference is statistically significant for the overall and elementary grades and marginally significant for middle school. This represents a large, meaningful difference: in our sample, roughly 26% of teachers depart their school in any given year, making a difference of 7 percentage points equal to a 27% reduction in turnover.

Columns 4-6 present impacts on teacher evaluation ratings, which range from 1 to 4. Overall, we find that teamed teachers have significantly higher evaluation ratings than their non-teamed peers: scores are 0.103 points higher, on average ($p < .01$), and this effect seems to be the strongest in elementary schools.

Taken together, we find that NEW is associated with both lower turnover and higher evaluation ratings among teamed teachers. These associations do not, however, map neatly onto our achievement results: turnover reductions are broadly similar across school levels, and only marginally significant in middle school. The evaluation gains are concentrated in elementary grades, where we find no positive achievement effects. Even in middle school, the magnitudes of the turnover and evaluation effects are unlikely to explain the achievement results. We therefore view these as parallel teacher-level outcomes which cannot fully explain the mechanisms behind the middle school math effect.

7.6 Reviewing Mechanisms

Across these analyses, the mechanism evidence narrows the field more than it settles the question. Instructional differentiation seems to be the most plausible channel and receives some direct, if imprecise, support: within middle school math, the teaming effect rises with a school-grade's incoming achievement dispersion, as a differentiation explanation would predict, though the interaction is not statistically significant. We also find suggestive evidence that team composition differs across school levels, though evidence there is more limited.

The remaining analyses work largely by exclusion. Teaming tenure is distributed similarly across grades and subjects, so the effect is not an artifact of middle school math teams having been in place longer. Attendance shows no effect and cannot explain any achievement findings. Finally, the teacher-workforce channels—retention and evaluation ratings—are broadly similar across school levels, and thus do not follow the pattern of student achievement effects.

Overall, we cannot yet isolate the operative mechanism with confidence. Why the effect appears in middle school math and not elsewhere remains largely unresolved, and better understanding how differentiation and team composition drive results remains an important direction for future work.

8. Discussion

We provide the first plausibly causal evidence on the Next Education Workforce, a team-based staffing model designed to support teacher collaboration. We exploit the staggered rollout of the initiative across schools in a difference-in-differences model comparing teamed and non-teamed grades over time. We find a positive impact of NEW on middle school math achievement of about 0.08 SD. These middle school math effects are largely robust to alternate specifications, including models with teacher fixed effects and a suite of alternate heterogeneity-robust estimators.

Outside of middle school math, NEW does not appear to have produced positive effects. Estimates at the elementary level and for ELA are statistically insignificant or negative across some specifications. We interpret these findings cautiously, however, as the pre-trend violations documented above and the inconsistency of estimates across specifications limit the credibility of any strong causal claim in these subgroups. We also find that NEW is associated with significant reductions in teacher turnover rates and improvements in evaluation ratings, but these patterns do not align with the achievement effects and thus do not appear to be a plausible explanation for the gains concentrated in middle school math.

The middle school math effects of NEW are notable. Middle school is a critical juncture for mathematics, a period when many students' achievement and engagement in math stall or decline (Anderman, 2003; Marks, 2000). The magnitude is also moderately large: 0.08 SD is equivalent to about 53% of a teacher standard deviation, or the equivalent of going from a median teacher to a teacher at roughly the 70th percentile of the teacher effectiveness distribution (Hanushek & Rivkin, 2012). Put differently, the effect is comparable to the achievement improvement associated with a teacher's first year of classroom experience.

The positive middle school math effect aligns with the departmentalization literature. In elementary grades, where instruction is typically self-contained, NEW likely introduces specialization alongside coordination costs and weaker teacher-student relationships that Fryer (2018) argues can offset its benefits. Middle schools are more likely to already be departmentalized, so NEW can add the benefits of collaboration and shared responsibility without incurring costs of additional specialization. As such, middle schools are where we might expect, and ultimately find, convincing achievement gains. The concentration of effects in math rather than ELA is also consistent with previous literature on education interventions.

Our mechanism evidence, while limited, suggests the achievement gains do not operate primarily through the teacher-workforce channels often emphasized by proponents of strategic staffing. Although teamed teachers exhibit lower turnover and higher evaluation ratings, these patterns are similar across school levels and do not follow the pattern of achievement effects concentrated in middle school math. The evidence points instead toward improvements in instruction: the effect is most pronounced where incoming achievement is most dispersed, consistent with teaming operating through instructional differentiation rather than through improvements in teacher retention or effectiveness, though this analysis is descriptive and imprecise. This distinction matters for how the model is understood and scaled. If the achievement returns to teaming operate through instructional differentiation rather than through improved working conditions, then teaming is most likely to benefit students when implemented as a tool for instructional change rather than improved retention.

We note several key limitations of the work. First, we present suggestive evidence of compositional changes in grades and subjects implementing NEW, which are often paired with downward pre-trends in the years before implementation. While we use modern difference-in-differences estimators to account for bias from the canonical TWFE, not accounting for previous trends in the outcome could potentially be introducing omitted variable bias (Botosaru & Liu, 2025). Adequately accounting for these trends is at the current frontier of the methodological literature and a potential direction for future work as the field develops.

Second, our post-implementation panel is still relatively short, particularly for later adopters, limiting our ability to assess longer-run trajectories of student achievement. Given evidence of fade-out in interventions of this type (Backes et al., 2025), examining the durability of NEW's effects as additional cohorts accumulate will be an important priority for future work. Finally, our data shed little light on what is actually happening inside classrooms—how teams function, the depth of teaming students actually experience, the extent to which teachers adopt NEW's practices, and whether and how those practices reshape instruction and learning. Future research should open this black box, connecting implementation fidelity to student outcomes in ways the present study cannot.

References

Anderman, L. H. (2003). Academic and social perceptions as predictors of change in middle school students' sense of school belonging. *The Journal of Experimental Education, 72*(1), 5–22. https://doi.org/10.1080/00220970309600877

Atteberry, A., Loeb, S., & Wyckoff, J. (2017). Teacher churning: Reassignment rates and implications for student achievement. *Educational Evaluation and Policy Analysis, 39*(1), 3–30. https://doi.org/10.3102/0162373716659929

Bacher-Hicks, A., Chin, M., Kane, T.J., & Staiger, D.O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review, 73*, 101919. https://doi.org/10.1016/j.econedurev.2019.101919

Backes, B., Cowan, J., & Goldhaber, D. (2026). Departmentalized instruction and elementary school effectiveness. *Journal of Public Economics, 261*, 105715. https://doi.org/10.1016/j.jpubeco.2026.105715

Backes, B., Cowan, J., & Goldhaber, D. (2025). Building bridges to middle school? Elementary school departmentalization and academic achievement in the upper grades (CALDER Working Paper No. 315-0325). *National Center for Analysis of Longitudinal Data in Education Research (CALDER)*. https://caldercenter.org/sites/default/files/2025-04/CALDER-WP-315-0325_0.pdf

Backes, B., & Hansen, M. (2018). Reaching further and learning more? Evaluating Public Impact's Opportunity Culture initiative (CALDER Working Paper No. 181). *National Center for Analysis of Longitudinal Data in Education Research (CALDER)*. https://caldercenter.org/publications/reaching-further-and-learning-more-evaluating-public-impacts-opportunity-culture

Basile, C., Maddin, B., & Audrain, L. (2023). *The Next Education Workforce: How team-based staffing models can support equity and improve learning outcomes.* Rowman & Littlefield.

Bastian, K. C., & Fortner, C. K. (2020). Is less more? Subject-area specialization and outcomes in elementary schools. *Education Finance and Policy, 15*(2), 357–382. https://doi.org/10.1162/edfp_a_00278

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies, 91*(6), 3253–3285. https://doi.org/10.1093/restud/rdae007

Botosaru, I., & Liu, L. (2025). Time-varying heterogeneous treatment effects in event studies. Working paper. arXiv. https://doi.org/10.48550/arXiv.2509.13698

Callaway, B., & Sant'Anna, P. H.C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics, 134*(3), 1405–1454. https://doi.org/10.1093/qje/qjz014

COVID-19 School Data Hub. (2022). Arizona school learning model data, 2020–2021. Arizona Department of Education. https://www.covidschooldatahub.com/states/arizona

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review, 110*(9), 2964–2996. https://doi.org/10.1257/aer.20181169

Dee, J. R., Henkin, A. B., & Singleton, C. A. (2006). Organizational commitment of teachers in urban schools: Examining the effects of team structures. *Urban Education, 41*(6), 603–627. https://doi.org/10.1177/0042085906292512

Dietrichson, J., Filges, T., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Jensen, U. H. (2020). Targeted school-based interventions for improving reading and mathematics for students with, or at risk of, academic difficulties in Grades 7–12: A systematic review. *Campbell Systematic Reviews: Better Evidence for a Better World, 16*(2). https://doi.org/10.1002/cl2.1081

Fryer, R. G. (2018). The "pupil" factory: Specialization and the production of human capital in schools. *American Economic Review, 108*(3), 616–656. https://doi.org/10.1257/aer.20161495

Gardner, J. (2022). Two-stage differences in differences. Working paper. https://jrgcmu.github.io/2sdd_current.pdf

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics, 4*, 131–157. https://doi.org/10.1146/annurev-economics-080511-111001

Hemelt, S. W., Ladd, H. F., & Clifton, C. R. (2021). Do teacher assistants improve student outcomes? Evidence from school funding cutbacks in North Carolina. *Educational Evaluation and Policy Analysis, 43*(2), 280–304. https://doi.org/10.3102/0162373721990361

Hwang, N., & Kisida, B. (2022). Spread too thin: The effect of specialization on teaching effectiveness. *Educational Evaluation and Policy Analysis, 44*(4), 593–607. https://doi.org/10.3102/01623737221084312

Ingersoll, R., Audrain, L., & Laski, M. (2025). Team-based staffing, teacher authority, and teacher turnover. Center on Reinventing Public Education. https://crpe.org/team-based-staffing-teacher-authority-and-teacher-turnover/

Jones, N., & Winters, M. A. (2024). Are two teachers better than one?: The effect of coteaching on students with and without disabilities. *Journal of Human Resources, 59*(4), 1180–1206. https://doi.org/10.3368/jhr.0420-10834R3

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* MET Project, Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation. https://files.eric.ed.gov/fulltext/ED540960.pdf

Kirksey, J. J., Gottlieb, J. J., Wiseman, A. W., Reed, B. M., Lansford, T., & Robison, E. (2026). *Evaluating Opportunity Culture: Improving student achievement through strategic staffing in West Texas* (Policy Brief No. 5). Texas Tech University, Center for Innovative Research in Change, Leadership, and Education. https://hdl.handle.net/2346/104883

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kraft, M. A., & Lyon, M. A. (2024). The rise and fall of the teaching profession: Prestige, interest, preparation, and satisfaction over the last half century. *American Educational Research Journal, 61*(6), 1192–1236. https://doi.org/10.3102/00028312241276856

Laski, M. (2024). Early evidence of improved educator outcomes in Next Education Workforce™ models. Center on Reinventing Public Education. https://crpe.org/early-evidence-of-improved-educator-outcomes-in-next-education-workforcetm-models/

Laski, M., Rainey, L., & Kightlinger, B. (2026). *Review of research on characteristics of reimagined teaching roles.* The Center for Inspired Teaching & Exceptional Learning (CITEL). https://citel.teachplus.org/

Leithwood, K., Harris, A., & Hopkins, D. (2020). Seven strong claims about successful school leadership revisited. *School Leadership & Management, 40*(1), 5–22. https://doi.org/10.1080/13632434.2019.1596077

Lortie, D. C. (1975). *Schoolteacher: A sociological study*. University of Chicago Press.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153–184. https://doi.org/10.3102/00028312037001153

Next Education Workforce. (n.d.). How teams work. Arizona State University. https://workforce.education.asu.edu/how-teams-work/

Public Impact. (n.d.). The Opportunity Culture® strategy. https://www.opportunityculture.org/the-strategy/

Rockoff, J. E., & Lockwood, B. B. (2010). Stuck in the middle: Impacts of grade configuration in public schools. *Journal of Public Economics, 94*(11–12), 1051–1061. https://doi.org/10.1016/j.jpubeco.2010.06.017

Ronfeldt, M., Farmer, S. O., McQueen, K., & Grissom, J. A. (2015). Teacher collaboration in instructional teams and student achievement. *American Educational Research Journal, 52*(3), 475–514. https://doi.org/10.3102/0002831215585562

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal, 50*(1), 4–36. https://doi.org/10.3102/0002831212463813

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights 4*(3), 305–322. https://doi.org/10.1257/aeri.20210236

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics, 235*, 2218–2244. https://doi.org/10.1016/j.jeconom.2023.03.008

Sant'Anna, P. H. C., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics, 219*(1), 101–122. https://doi.org/10.1016/j.jeconom.2020.06.003

U.S. Department of Education. (2026). *Guidance: Strategic staffing under Title II, Part A.* https://www.ed.gov/media/document/guidance-strategic-staffing-under-title-ii-part-113188.pdf

Tables and Figures

Table 1. Implementation Across Contexts

A. Proportion of schools with teams

Subject	School level	2021	2022	2023	2024	2025
ELA	Elementary	0.13	0.10	0.18	0.34	0.14
ELA	Middle	0.09	0.17	0.18	0.38	0.35
Math	Elementary	0.12	0.10	0.17	0.34	0.13
Math	Middle	0.11	0.16	0.10	0.32	0.28

B. Proportion of students in teamed grades assigned directly to teamed teachers

Subject	School level	2021	2022	2023	2024	2025
ELA	Elementary	0.79	0.92	0.88	0.89	0.87
ELA	Middle	1.00	0.66	0.65	0.67	0.74
Math	Elementary	0.84	0.92	0.95	0.89	0.95
Math	Middle	1.00	0.77	0.96	0.68	0.97

Note: Table includes school-subject units for school years 2021-2025, by school level. Panel A displays the proportion of units with at least one teamed teacher in that subject-grade-level. Panel B displays the proportion of students assigned directly to a teamed teacher within units with at least one teamed teacher.

Table 2. Team Composition

Year	Level	Number: Teams	Number: Team members	Number: Unique roles	Proportion teams with… Lead teacher	Proportion teams with… Para-educator	Proportion teams with… Special education teacher	Proportion teams with… Teacher candidate
2023	Elem	3.24	4.19	2.39	0.64	0.09	0.33	0.31
2023	Middle	4.33	4.13	2.50	0.75	0.00	0.50	0.08
2024	Elem	4.19	3.60	1.67	0.39	0.11	0.15	0.04
2024	Middle	4.00	3.83	1.92	0.33	0.00	0.50	0.17

Note: Team identifiers collected by NEW staff at ASU and merged into administrative data by district staff. Data are at the individual level and include all teamed teachers. Each cell presents school-level averages by school level and year.

Table 3. Descriptive Statistics by Treated Status

	Ever treated	Never treated	Difference
Students	(1)	(2)	(3)
American Indian/Alaskan Native	0.073	0.058	0.014**
Asian	0.011	0.012	-0.001
Black/African American	0.075	0.070	0.005
Hispanic	0.511	0.465	0.046***
Native Hawaiian/Pacific Islander	0.013	0.011	0.002
White	0.774	0.791	-0.017*
Two or more races	0.054	0.058	-0.004
Female	0.485	0.480	0.005
Special Education	0.218	0.214	0.004
English Language Learner	0.155	0.126	0.030***
ELA test score (SD)	-0.106	0.038	-0.145***
Math test score (SD)	-0.149	0.026	-0.175***
ELA proficiency	0.195	0.264	-0.069***
Math proficiency	0.148	0.236	-0.088***
Attendance rate	0.907	0.911	-0.004
Chronic absenteeism	0.317	0.295	0.021*
Teachers			
Teacher experience	11.141	12.522	-1.381**
Novice	0.262	0.179	0.083***
Standard license	0.955	0.936	0.019*
N observations	*239*	*1057*	

Note: The unit of observation is school-grade-year. The sample includes school-grade-years (grades 3-8) from 2021 to 2025. Student characteristics reflect school-grade-year means. Test scores are from the year before implementation and are standardized at the subject-year level. Chronic absenteeism is defined as having an attendance rate lower than 90%. Novice is defined as three or fewer years of teaching experience. Column 1 reflects school-grade-years ever assigned to the NEW teaming model. Differences in column 3 are from two-sample t-tests assuming unequal variances.

+p < .1; *p < .05; **p < .01; ***p < .001

Table 4. Trends in Covariates Prior to NEW Adoption

	Math Elem	Math Middle	ELA Elem	ELA Middle
	(1)	(2)	(3)	(4)
Standardized scale score	-0.020**	-0.005	-0.007	-0.018
Total students	-0.897**	-7.117***	-0.810**	-3.428
Hispanic	0.049	-0.757	0.001	-0.680
American Indian/Alaskan Native	-0.259**	-0.265	-0.269**	-0.567
Black/African American	0.016	0.036	0.020	0.638
White	0.166	0.222	0.190	0.192
Special Education	0.087	0.109	0.166	-0.404
English Language Learner	0.047	0.501	0.006	0.771***
Average daily attendance	-0.087*	-0.237	-0.082	-0.346**
Teacher experience	-0.580**	0.377	-0.255	0.752
N observations	*1778*	*226*	*1778*	*236*

Note: Each cell presents output from a separate regression where the coefficient of interest is on a linear trend in the number of years prior to NEW adoption. The unit of observation is school-grade-year. Student characteristics reflect school-grade-year means. Regressions include school-by-grade and year fixed effects and standard errors are clustered at the school-by-grade level. Elem = 3-6, Middle = 7-8.
+p < .1; *p < .05; **p < .01; ***p < .001

Table 5. Relationships Between Covariates and NEW Adoption Year

	Math Elem	Math Middle	ELA Elem	ELA Middle
	(1)	(2)	(3)	(4)
Standardized scale score†	0.736*	1.573	0.415	1.407**
Total students†	-0.006	0.003	-0.005	0.001
American Indian/Alaskan Native	0.032**	-0.047***	0.034***	-0.021**
Black/African American	0.029	-0.012	0.039	0.002
White†	-0.018*	0.021	-0.021*	0.017*
Hispanic	-0.004	-0.006	-0.003	0.010
Special Education†	-0.033	-0.002	-0.029	0.090***
English Language Learner†	-0.023	-0.053	-0.022	-0.072
Average daily attendance	-0.092	0.051	-0.088	0.036**
Teacher experience†	0.014	0.061	0.012	-0.155*
F-test p-value: full slate of covariates	0.021	0.346	0.055	0.129
Residual DF	82	1	83	6
F-test p-value: selected covariates(†)	0.163	0.059	0.350	0.142
Residual DF	86	5	87	10
N observations	*93*	*12*	*94*	*17*

Note: Each cell presents output from a separate regression estimating the relationship between a school characteristic in our baseline year (2017) and the year of NEW adoption. The unit of observation is school-grade. Student characteristics reflect school-grade means in 2017. All regressions include heteroskedasticity-robust standard errors. F-test p-value tests the null hypothesis that characteristics do not jointly predict adoption timing. Elem = 3-6, Middle = 7-8. +p < .1; *p < .05; **p < .01; ***p < .001

Table 6. Relationships Between Students' Exposure to a Teamed Teacher and Standardized Test Achievement

	Math			ELA		
	Overall	Elem	Middle	Overall	Elem	Middle
	(1)	(2)	(3)	(4)	(5)	(6)
Canonical TWFE	-0.005	-0.043+	0.075+	-0.004	-0.019	0.021
	(0.022)	(0.024)	(0.044)	(0.014)	(0.017)	(0.026)
BJS	-0.020	-0.073*	0.100*	-0.024	-0.048*	0.014
	(0.029)	(0.030)	(0.047)	(0.016)	(0.020)	(0.029)
	[0.108]	[0.036]*	[0.836]	[0.035]*	[0.773]	[0.006]**
dCDH	0.033	0.024	0.056	-0.010	0.030	-0.030
	(0.034)	(0.047)	(0.057)	(0.026)	(0.039)	(0.039)
	[0.538]	[0.641]	[0.620]	[0.163]	[0.148]	[0.032]*
N observations	*187247*	*128038*	*59209*	*195171*	*129386*	*65785*
N school-grades	*261*	*229*	*32*	*265*	*229*	*36*

Note: Each cell represents output from a separate regression using data at the student-year level. The outcome is identified in each column header. The first panel uses the canonical TWFE model from equation (1). The second panel uses the BJS estimator using the did_imputation command in Stata. The third panel uses the dCDH estimator using the did_multiplegt_dyn command in Stata. Achievement is standardized within year-grade-subject cells using the district-wide sample. The independent variable is a teamed indicator that uses the inclusive definition of teaming within a given year and is equal to 1 for all students in a school-grade-subject-year cell where there is at least one teamed teacher. All regressions include school-by-grade and year fixed effects. Standard errors are clustered at the school-by-grade level and presented in parentheses. P-values on BJS pre-trend tests and dCDH joint nullity of placebos are in brackets. Overall = grades 3-8, Elem = 3-6, Middle = 7-8.
+p < .1; *p < .05; **p < .01; ***p < .001

Table 7. Test Score Model Specifications With Teacher Fixed Effects

	Math			ELA			Attendance		
	Overall	Elem	Middle	Overall	Elem	Middle	Overall	Elem	Middle
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Teamed effect	0.027	0.014	0.067+	0.009	0.024	-0.022	-0.003	-0.001	-0.006*
	(0.017)	(0.019)	(0.035)	(0.020)	(0.017)	(0.049)	(0.002)	(0.001)	(0.003)
N observations	*208076*	*139617*	*68459*	*222446*	*147080*	*75364*	*374489*	*197288*	*177198*
N school-grades	*265*	*230*	*35*	*268*	*230*	*38*	*271*	*231*	*40*

Note: Each cell represents output from a separate regression model using data at the student-year-level. The outcome is identified in each column header. Achievement is standardized within year-grade-subject cells using the district-wide sample. In columns 7-9, the outcome is the proportion of days a student attended in a given year (0-1). The independent variable is a teamed indicator that uses the inclusive definition of teaming within a given year and is equal to 1 for all students in a school-grade-year cell where there is at least one teamed teacher. All regressions include teacher, school, grade, and year fixed effects. Standard errors are clustered at the school-by-grade level and presented in parentheses. Overall = grades 3-8, Elem = 3-6, Middle = 7-8.
+p < .1; *p < .05; **p < .01; ***p < .001

Table 8. Relationships Between Student and Teacher Characteristics and Teamed Impact

	Math Overall	Math Elem	Math Middle	ELA Overall	ELA Elem	ELA Middle
	(1)	(2)	(3)	(4)	(5)	(6)
SPED = 0	-0.028	-0.081	0.093	-0.034	-0.053	-0.005
SPED = 1	-0.010	-0.058	0.113	0.013	-0.036	0.079
Difference	0.018	0.023	0.020	0.047	0.016	0.084
	(0.032)	(0.034)	(0.037)	(0.029)	(0.031)	(0.059)
N untreated	*22590*	*15122*	*7468*	*26529*	*15512*	*11017*
N treated	*5691*	*4153*	*1538*	*6670*	*4195*	*2475*
Specification	1	1	2	1	1	1
ELL = 0	-0.029	-0.084	0.100	-0.016	-0.045	0.027
ELL = 1	0.049	0.029	0.091	-0.020	-0.024	0.042
Difference	0.078*	0.113**	-0.010	-0.004	0.021	0.015
	(0.031)	(0.043)	(0.045)	(0.032)	(0.037)	(0.083)
N untreated	*23512*	*16227*	*7285*	*28123*	*16666*	*11457*
N treated	*4767*	*3048*	*1719*	*5076*	*3041*	*2035*
Specification	1	1	2	1	1	1
Hispanic = 0	-0.051	-0.100	0.074	-0.028	-0.045	-0.000
Hispanic = 1	0.012	-0.040	0.116	-0.019	-0.049	0.028
Difference	0.063*	0.060	0.042	0.009	-0.005	0.028
	(0.028)	(0.038)	(0.030)	(0.022)	(0.026)	(0.036)
N untreated	*13483*	*9731*	*3752*	*16079*	*10080*	*5999*
N treated	*14798*	*9544*	*5254*	*17120*	*9627*	*7493*
Specification	1	1	2	1	1	1
White = 0	-0.061	-0.126	0.092	-0.023	-0.082	0.109
White = 1	-0.013	-0.063	0.101	-0.025	-0.042	-0.004
Difference	0.049+	0.063+	0.008	-0.002	0.040	-0.113***
	(0.026)	(0.034)	(0.026)	(0.025)	(0.031)	(0.030)
N untreated	*6175*	*4106*	*2069*	*7003*	*4116*	*2887*
N treated	*22106*	*15169*	*6937*	*26196*	*15591*	*10605*
Specification	1	1	2	1	1	1
Novice = 0	-0.016	-0.030	0.020	-0.020	-0.019	-0.170
Novice = 1	-0.018	-0.108	0.146	-0.027	-0.071	0.004
Difference	-0.001	-0.078	0.125	-0.007	-0.052	0.174+
	(0.069)	(0.063)	(0.086)	(0.038)	(0.043)	(0.094)
N untreated	*12616*	*9241*	*3375*	*10460*	*9514*	*946*
N treated	*15391*	*9750*	*5641*	*22463*	*9908*	*12555*

Specification	1	1	2	1	1	2

Note: Each column in each panel represents output from a separate regression using data at the student-year level. The outcome is identified in each column header. Coefficients for indicator = 0 and indicator = 1 use the BJS estimator using the did_imputation command and the hetby() option in Stata. Achievement is standardized within year-grade-subject cells using the district-wide sample. The independent variable is a teamed indicator that uses the inclusive definition of teaming within a given year and is equal to 1 for all students in a school-grade-subject-year cell where there is at least one teamed teacher. All regressions include school-by-grade-by-indicator and year-by-indicator fixed effects. Standard errors are clustered at the school-by-grade level and presented in parentheses. Specification 1 uses the leave-out procedure. Specification 2 reports standard errors from a more conservative pooled-variance estimator, used in cells where the leave-out correction is not computable because of the small number of treated clusters. Overall = grades 3-8, Elem = 3-6, Middle = 7-8.
+p < .1; *p < .05; **p < .01; ***p < .001

Table 9. Relationships Between Students' Exposure to a Teamed Teacher and Attendance

	Attendance Overall	Attendance Elem	Attendance Middle
	(1)	(2)	(3)
Canonical TWFE	-0.006*	-0.002+	-0.006+
	(0.002)	(0.001)	(0.003)
BJS	-0.007***	-0.004**	-0.010***
	(0.002)	(0.001)	(0.002)
	[0.000]***	[0.284]	[0.000]***
dCDH	-0.001	-0.003	-0.000
	(0.003)	(0.002)	(0.002)
	[0.000]***	[0.647]	[0.000]***
N observations	*239041*	*155235*	*83806*
N school-grades	*267*	*230*	*37*

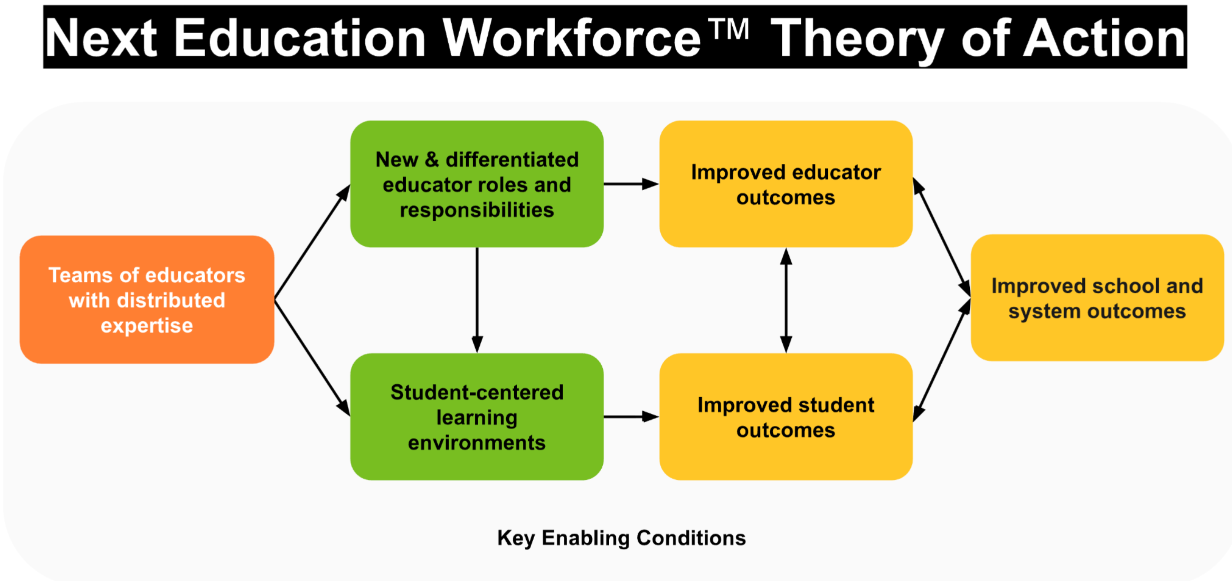
Note: Each cell represents output from a separate regression using data at the student-year level. The outcome is identified in each column header. The first panel uses the canonical TWFE model from equation (1). The second panel uses the BJS estimator using the did_imputation command in Stata. The third panel uses the dCDH estimator using the did_multiplegt_dyn command in Stata. The outcome is the proportion of days a student attended in a given year (0-1). The independent variable is a teamed indicator that uses the inclusive definition of teaming within a given year and is equal to 1 for all students in a school-grade-year cell where there is at least one teamed teacher. All regressions include school-by-grade and year fixed effects. Standard errors are clustered at the school-by-grade level and presented in parentheses. P-values on BJS pre-trend tests and dCDH joint nullity of placebos are in brackets. Overall = grades 3-8, Elem = 3-6, Middle = 7-8.
+p < .1; *p < .05; **p < .01; ***p < .001

Table 10. Relationships Between Teacher's Teamed Status and Teacher Turnover and Evaluation Ratings

	Turnover			Evaluation Ratings		
	Overall	Elem	Middle	Overall	Elem	Middle
	(1)	(2)	(3)	(4)	(5)	(6)
Teamed effect	-0.073**	-0.069*	-0.076+	0.103**	0.115*	0.077
	(0.023)	(0.027)	(0.039)	(0.038)	(0.048)	(0.058)
N observations	*8564*	*5766*	*2798*	*7726*	*5174*	*2552*
N school-grades	*263*	*229*	*34*	*262*	*227*	*35*

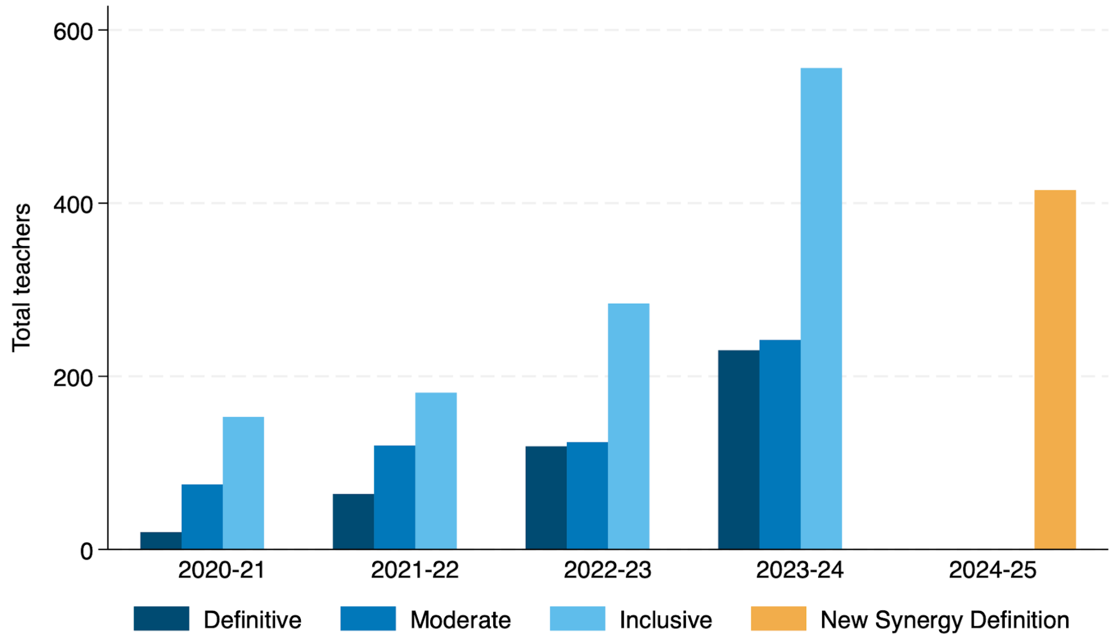
Note: Each cell represents output from a separate regression model using data at the teacher-year-level following equation (3). In columns 1-3, the outcome variable is a binary equal to 1 if a teacher departed their school in year *t+1*. In columns 4-6, the outcome variable is the teacher's evaluation rating, ranging from 1 to 4. The independent variable is a teamed indicator that uses the inclusive definition of teaming within a given year and is equal to 1 for all teachers identified as teaming in a given year. Standard errors are clustered at the school-by-grade level and are presented in parentheses. Overall = grades 3-8, Elem = 3-6, Middle = 7-8.
+p < .1; *p < .05; **p < .01; ***p < .001

Figure 1. Next Education WorkforceTM Theory of Action



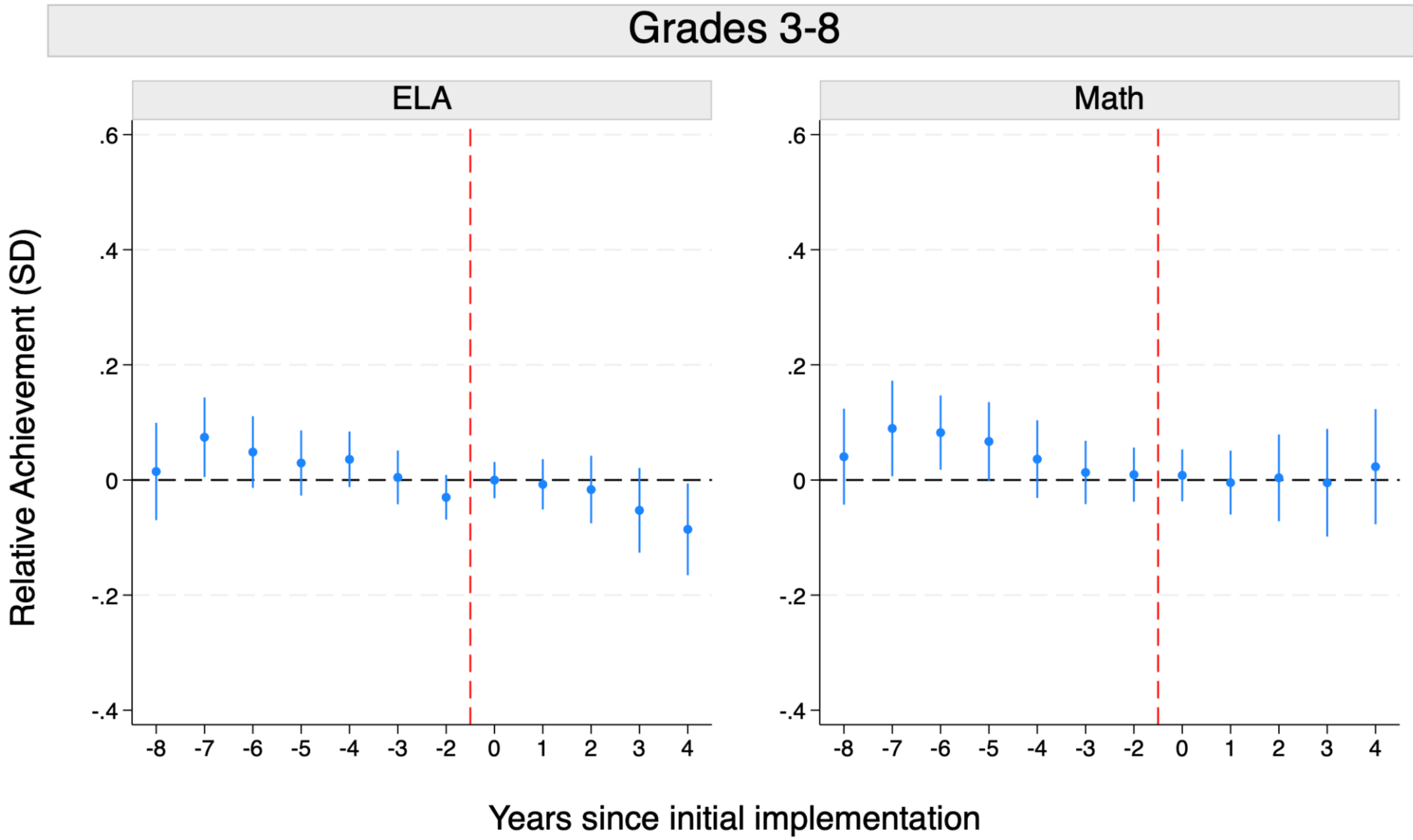
Source: Next Education Workforce website, https://workforce.education.asu.edu/results/#theory-of-action

Figure 2. Teacher-Level Teaming Rates, by Year and Definition



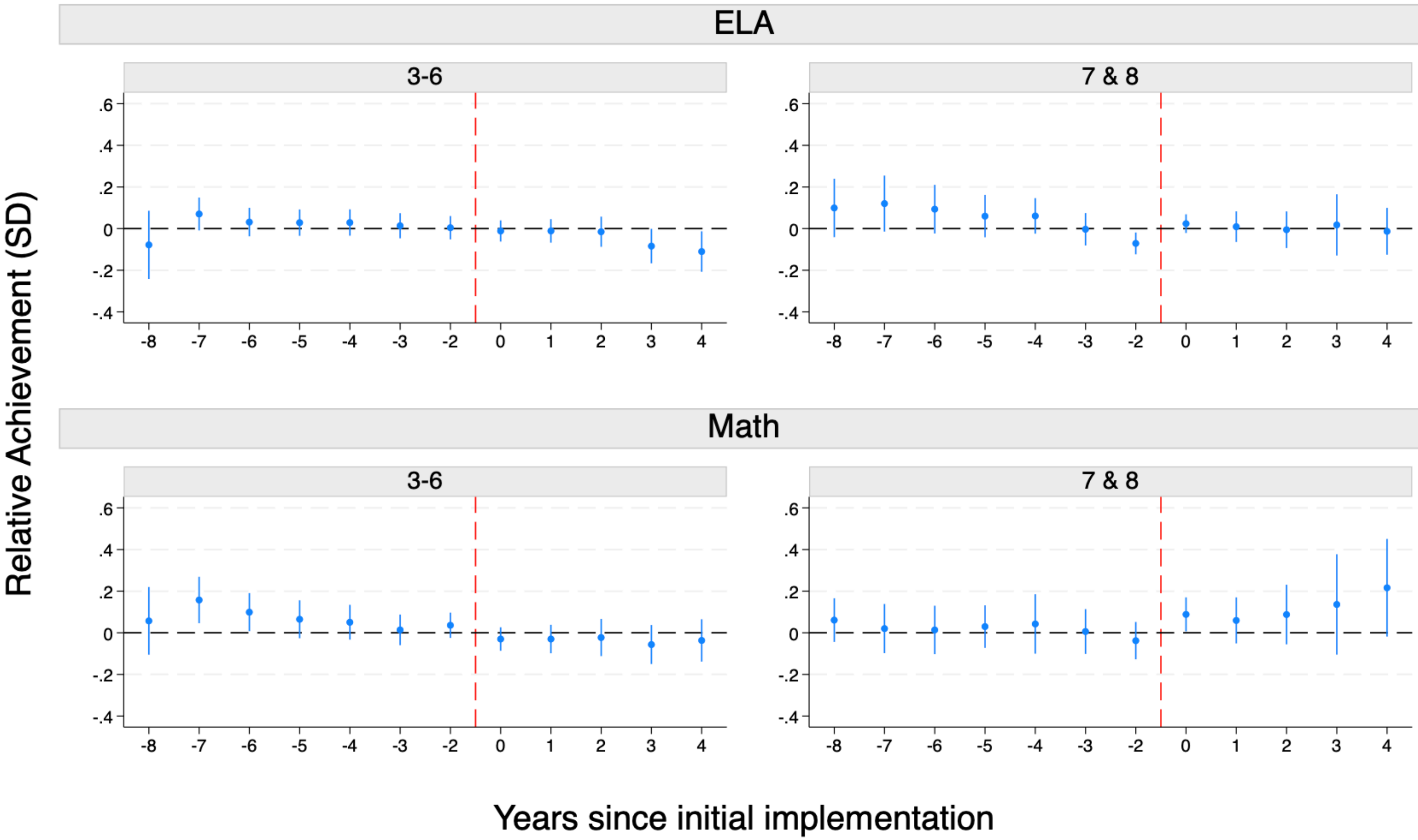
Note: We use the most inclusive teaming definition in our analyses.

Figure 3. Event Study Figures From Canonical TWFE Model, Grades 3-8 by Subject



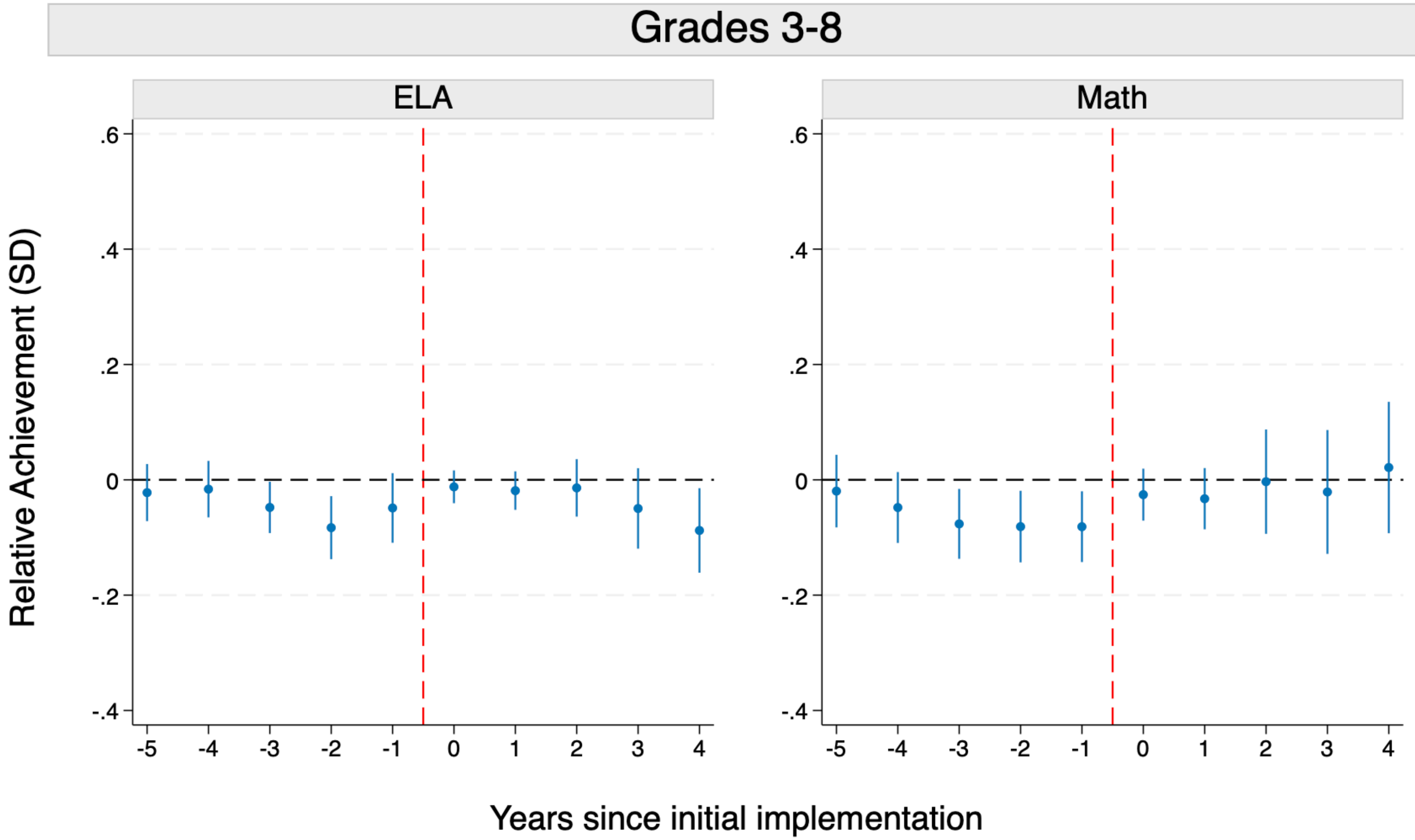
Note: Each dot represents a separate coefficient on a dummy variable for years since implementation. All non-implementers have a 0 for each dummy variable. Vertical lines represent 95% confidence intervals. Y-axis is relative achievement, in standard deviation units. Estimates come from the canonical TWFE model with school-by-grade and year fixed effects and no controls. Standard errors are clustered at the school-by-grade level.

Figure 4. Event Study Figures From Canonical TWFE Model, Subject by Grade-Level



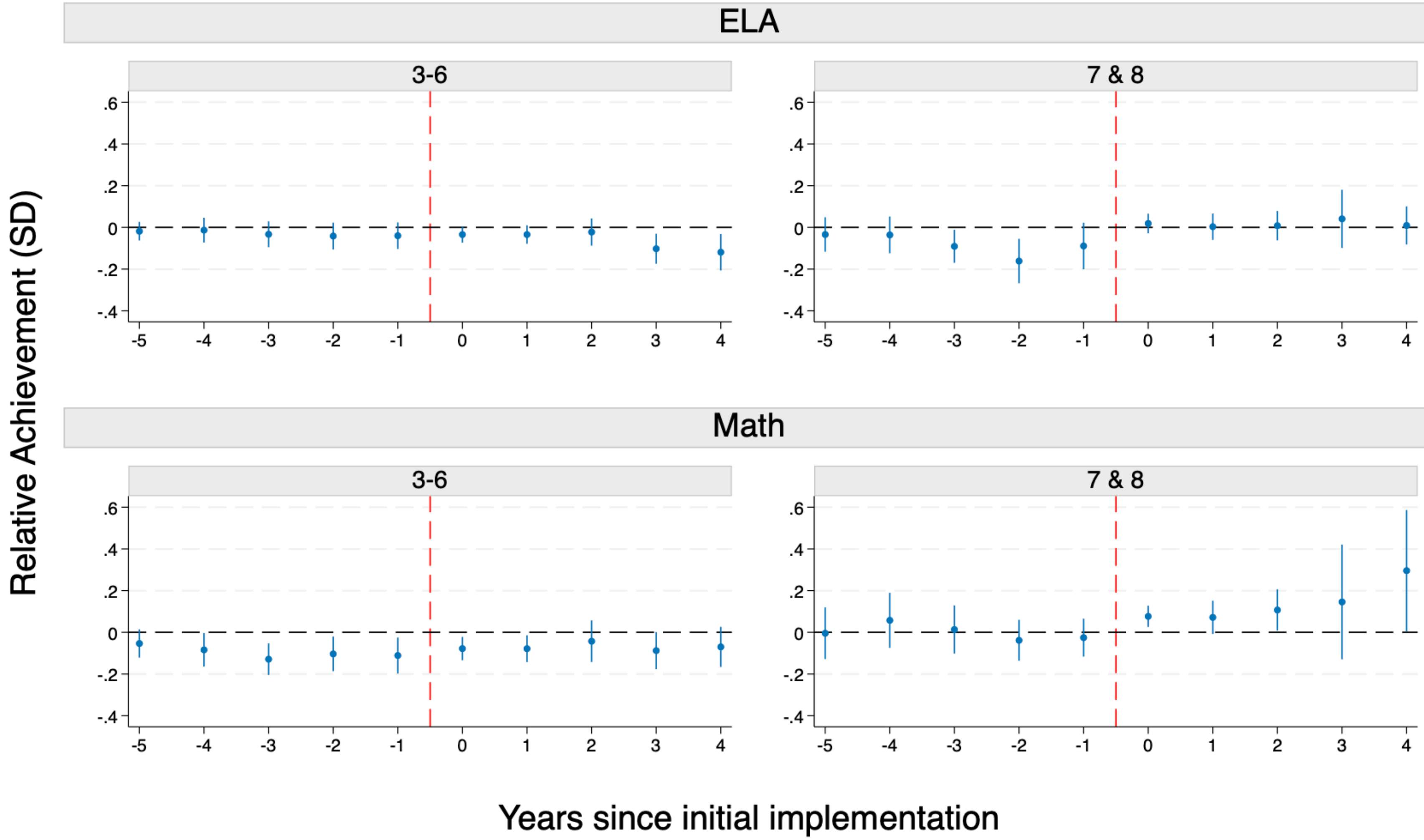
Note: Each dot represents a separate coefficient on a dummy variable for years since implementation. All non-implementers have a 0 for each dummy variable. Vertical lines represent 95% confidence intervals. Y-axis is relative achievement, in standard deviation units. Estimates come from the canonical TWFE model with school-by-grade and year fixed effects and no controls. Standard errors are clustered at the school-by-grade level.

Figure 5. Event Study Figures From BJS Estimator, Grades 3-8 by Subject



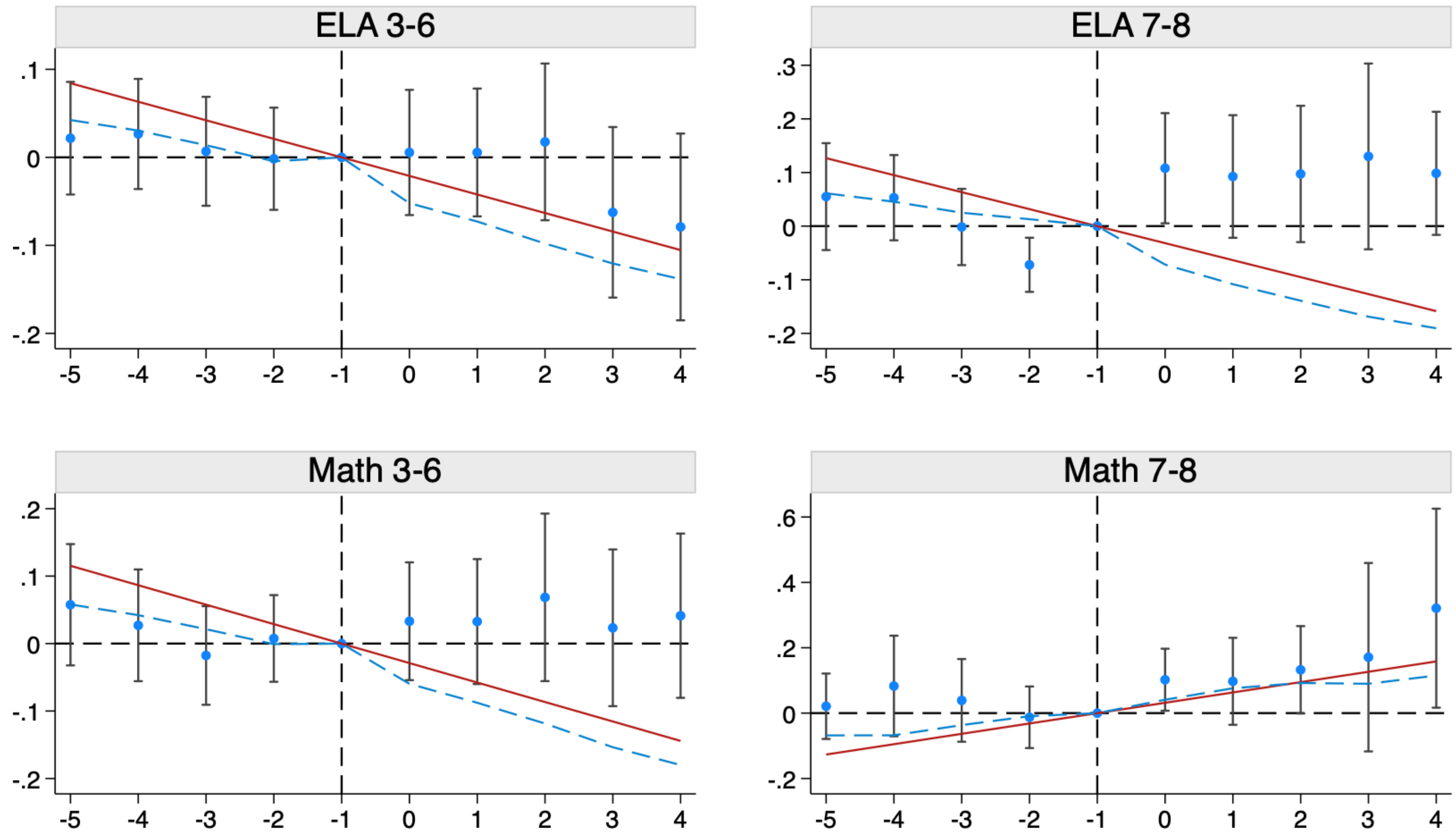
Note: Each dot represents a separate coefficient on a dummy variable for years since implementation. All non-implementers have a 0 for each dummy variable. Vertical lines represent 95% confidence intervals. Y-axis is relative achievement, in standard deviation units. Estimates come from the BJS estimator with school-by-grade and year fixed effects. Standard errors are clustered at the school-by-grade level. Joint test of pre-period parallel trends has a p-value of .035 for ELA and .108 for math.

Figure 6. Event Study Figures From BJS Estimator, Subject by Grade-Level



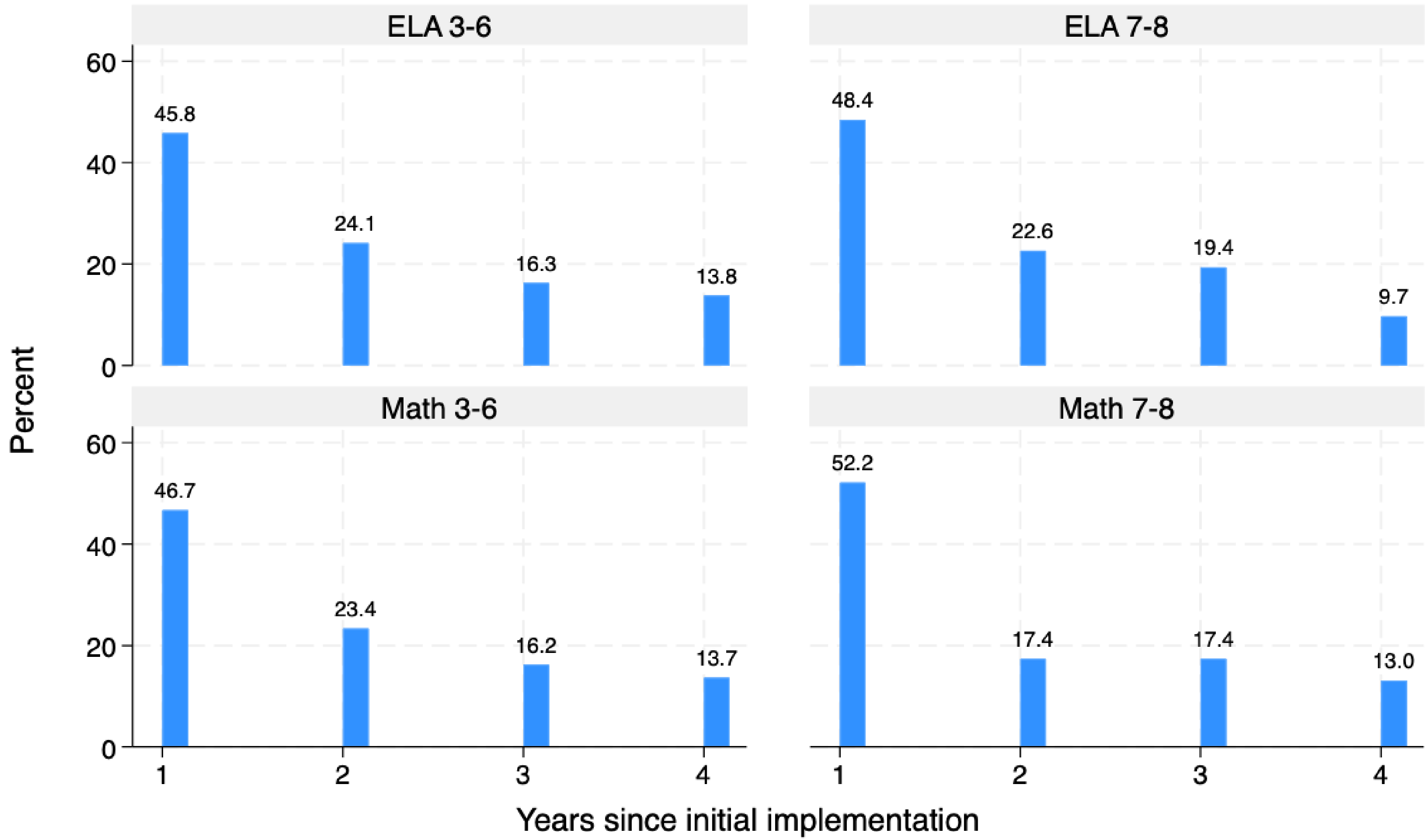
Note: Each dot represents a separate coefficient on a dummy variable for years since implementation. All non-implementers have a 0 for each dummy variable. Vertical lines represent 95% confidence intervals. Y-axis is relative achievement, in standard deviation units. Estimates come from the BJS estimator with school-by-grade and year fixed effects. Standard errors are clustered at the school-by-grade level. Joint test of pre-period parallel trends has a p-value of .773 for elementary ELA, .006 for middle school ELA, .036 for elementary math, and .836 for middle school math.

Figure 7. Testing Precision of Parallel Pre-Trends



Note: Each panel presents coefficients from a separate BJS estimator, estimated with five pre-periods and renormalized relative to year -1. The BJS estimator natively estimates all pre-period coefficients, while the Roth diagnostic requires year -1 as a reference year. This renormalization increases the standard errors by approximately 20% in math and 30% in ELA, yielding a conservative joint pre-trend test. Panels are presented for slope diagnostics only; renormalization shifts each coefficient by the value of the year -1 coefficient. Y-axis is relative achievement, in standard deviation units. The solid red line represents the smallest linear pre-period trend detectable with 80% power, passing directly through year -1 and oriented in the direction of the estimated treatment effect. The dashed blue line represents the path that all coefficients would take if this hypothesized linear trend existed within the data, given the mechanics of the BJS estimator. Standard errors clustered at the school-by-grade level.

Figure 8. Percent Teamed School-Grade-Years Across Years of Implementation



Note: Percent of teamed school-grade unit-years at each year since initial implementation, by school level and subject. A school-grade contributes once for each year it is teamed; bars sum to 100% within each panel.

Appendix Table 1. P-values From BJS Tests of Parallel Trends by Number of Pre-Periods

	Math			ELA		
K pre-periods	Overall	Elem	Middle	Overall	Elem	Middle
	(1)	(2)	(3)	(4)	(5)	(6)
1	0.099+	0.161	0.541	0.630	0.455	0.880
2	0.066+	0.298	0.406	0.004**	0.419	0.002**
3	0.049*	0.024*	0.603	0.008**	0.559	0.002**
4	0.060+	0.036*	0.745	0.019*	0.709	0.005**
5	0.108	0.036*	0.836	0.035*	0.773	0.006**
6	0.162	0.063+	0.867	0.032*	0.793	0.011*
7	0.223	0.095+	0.421	0.042*	0.616	0.017*

Note: Each cell presents the associated p-value from the BJS joint test for no pre-trends using the did_imputation command and associated pretrends(K) option in Stata.
+p < .1; *p < .05; **p < .01; ***p < .001

Appendix Table 2. Diagnostic Statistics From Roth (2022) Test

	Detectable slope (80%)	Observed slope	Observed t	Joint p (-1 ref)	Joint p (native BJS)	SE inflation
	(1)	(2)	(3)	(4)	(5)	(6)
ELA 3-6	0.021	-0.009	-0.924	0.807	0.773	1.310
ELA 7-8	0.032	-0.044	-3.048	0.001	0.006	1.309
Math 3-6	0.029	-0.019	-1.589	0.340	0.036	1.213
Math 7-8	0.032	-0.015	-0.688	0.736	0.836	1.223

Note: Each row presents statistics from Roth pre-trend diagnostics, by school level and subject, estimated with five pre-periods. Column 1 presents the smallest linear pre-period trend detectable with 80% power, while column 2 presents the linear slope of the renormalized coefficients seen in the event study plots in Figure 7. Column 3 is the observed slope divided by its standard error. Columns 4 and 5 present joint tests of whether the pre-period coefficients are zero for the renormalized and native results. Column 6 presents the average ratio of the renormalized standard error to the native standard error for each coefficient. Standard errors clustered at the school-by-grade level.

Appendix Table 3. Relationships Between Students' Exposure to a Teamed Teacher and Standardized Achievement, Alternate Estimators

	Math Overall	Math Elem	Math Middle	ELA Overall	ELA Elem	ELA Middle
	(1)	(2)	(3)	(4)	(5)	(6)
Gardner (2022)	-0.002	-0.055*	0.096	-0.004	-0.028+	0.025
	(0.032)	(0.026)	(0.065)	(0.015)	(0.017)	(0.028)
Stacked DiD	-0.024	-0.074*	0.089+	-0.034*	-0.042*	-0.036
	(0.026)	(0.030)	(0.053)	(0.017)	(0.020)	(0.040)
Callaway and Sant'Anna (2021), unweighted	0.006	0.007	0.008	-0.008	0.003	-0.083+
	(0.030)	(0.033)	(0.061)	(0.025)	(0.029)	(0.043)
Callaway and Sant'Anna (2021), weighted	0.020	0.013	0.043	-0.017	0.012	-0.042
	(0.026)	(0.032)	(0.050)	(0.020)	(0.028)	(0.039)
N observations	*187391*	*128118*	*59273*	*195325*	*129466*	*65859*
N school-grades	*267*	*231*	*36*	*271*	*231*	*40*

Note: Each cell represents output from a separate regression. The first panel uses the Gardner estimator using the did2s command in Stata. The second panel uses the canonical TWFE model in a stacked dataset. The third and fourth panels use the Callaway and Sant'Anna estimator using the csdid command in Stata. Achievement is standardized within year-grade-subject cells using the district-wide sample. The independent variable is a teamed indicator that uses the inclusive definition of teaming within a given year and is equal to 1 for all students in a school-grade-subject-year cell where there is at least one teamed teacher. Data for the Gardner (2022) estimator and stacked DiD are at the student-year level. The outcome is identified in each column header. Callaway and Sant'Anna (2021) uses data collapsed to the school-grade-year level. All regressions include school-by-grade and year fixed effects. Standard errors are clustered at the school-by-grade level and presented in parentheses. Overall = grades 3-8, Elem = 3-6, Middle = 7-8.

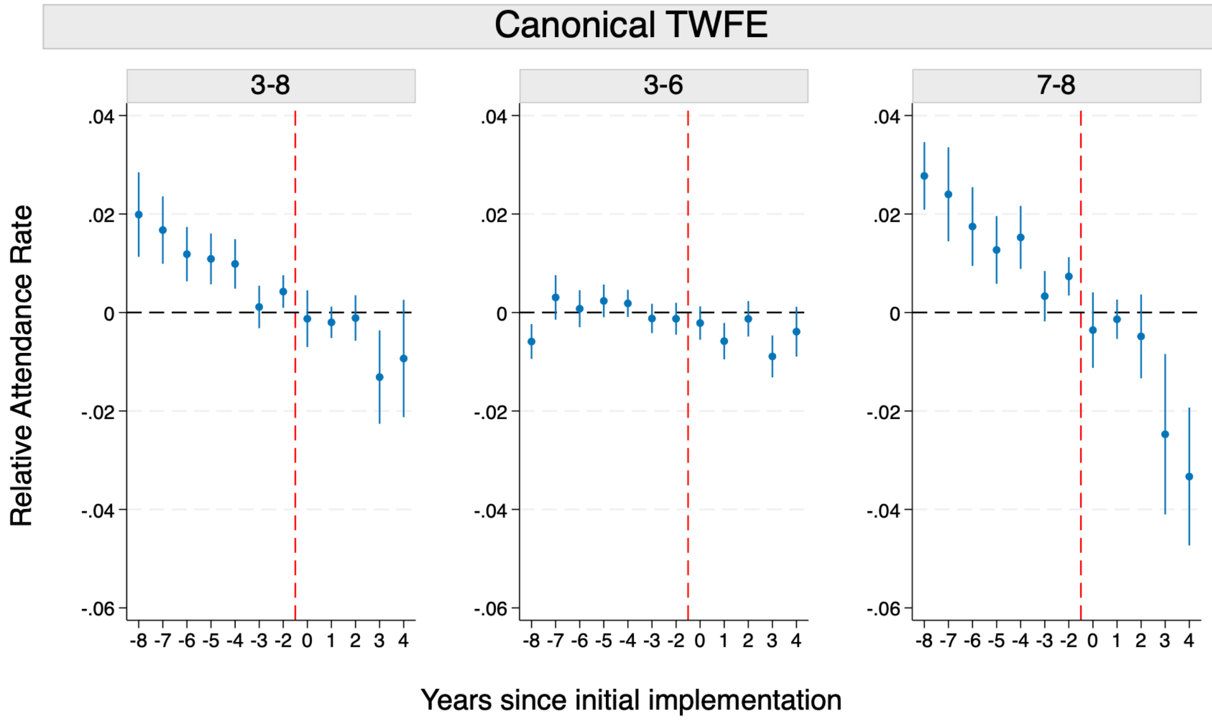
+p < .1; *p < .05; **p < .01; ***p < .001

Appendix Table 4. Teaching Assignments by School Level, Subject, and Treatment Status

Subject	Grades	Group	Stayer	Newcomer	Within-school	Other school	Previously in district	First-year teacher	Experienced, other district	N
			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ELA	3-6	Never-teamed	0.66	0.34	0.08	0.05	0.03	0.06	0.12	3270
ELA	3-6	Treated, pre-period	0.60	0.40	0.09	0.05	0.02	0.08	0.15	1397
ELA	3-6	Treated, year 1	0.58	0.42	0.15	0.04	0.03	0.06	0.13	301
ELA	7-8	Never-teamed	0.59	0.41	0.09	0.04	0.02	0.09	0.18	456
ELA	7-8	Treated, pre-period	0.64	0.36	0.06	0.03	0.01	0.07	0.18	665
ELA	7-8	Treated, year 1	0.60	0.40	0.10	0.09	0.04	0.03	0.14	108
ELA	All	Never-teamed	0.65	0.35	0.08	0.04	0.03	0.07	0.13	3726
ELA	All	Treated, pre-period	0.62	0.38	0.08	0.05	0.02	0.08	0.16	2062
ELA	All	Treated, year 1	0.59	0.41	0.14	0.06	0.03	0.05	0.13	409
Math	3-6	Never-teamed	0.66	0.34	0.09	0.05	0.03	0.06	0.12	3096
Math	3-6	Treated, pre-period	0.61	0.39	0.09	0.05	0.02	0.08	0.15	1319
Math	3-6	Treated, year 1	0.57	0.43	0.17	0.04	0.03	0.05	0.14	295
Math	7-8	Never-teamed	0.69	0.31	0.07	0.05	0.03	0.04	0.12	504
Math	7-8	Treated, pre-period	0.63	0.37	0.09	0.04	0.02	0.10	0.12	321
Math	7-8	Treated, year 1	0.55	0.45	0.09	0.05	0.15	0.02	0.15	55
Math	All	Never-teamed	0.66	0.34	0.09	0.05	0.03	0.06	0.12	3600
Math	All	Treated, pre-period	0.61	0.39	0.09	0.05	0.02	0.08	0.14	1640
Math	All	Treated, year 1	0.57	0.43	0.15	0.04	0.05	0.05	0.14	350

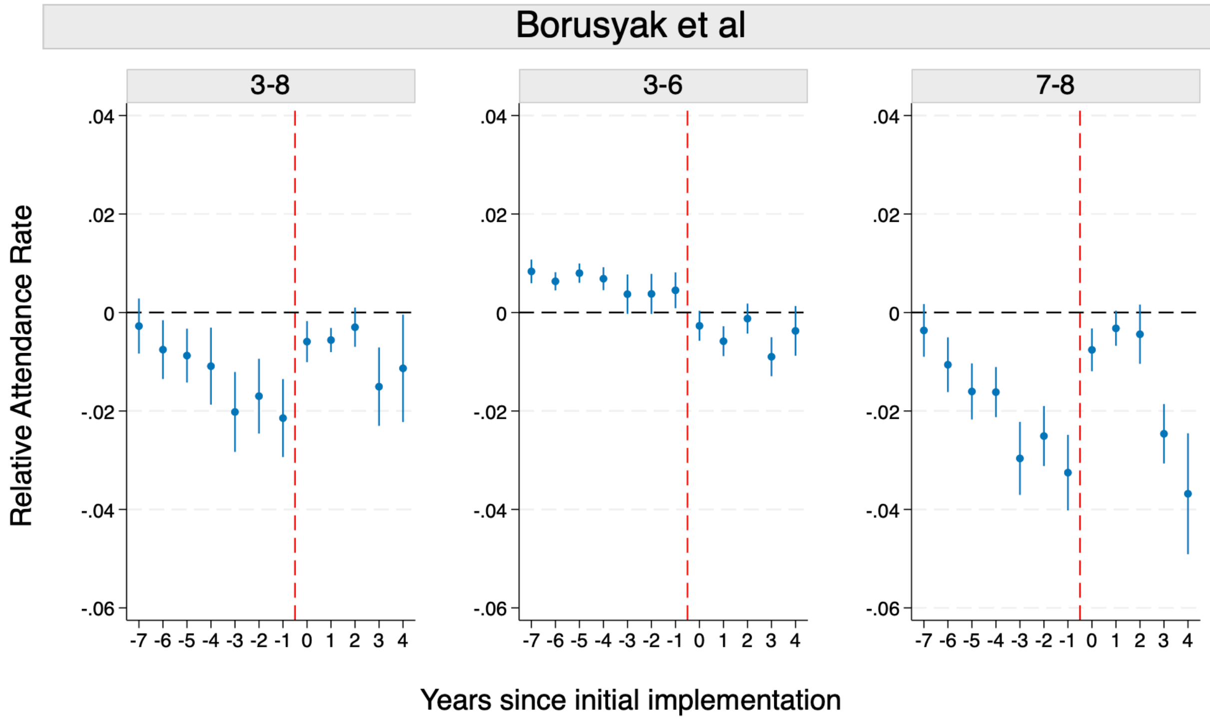
Note: Table describes prior teaching assignments for teachers within each school level and subject. For each year *t,* values give the proportion of teacher-school-grade-subject-year observations according to their assignment status in year *t-1*. The "Stayer" value is the proportion of teachers in the same teaching assignment between years. The "Newcomer" value represents teachers in different teaching assignments between years. "Newcomers" can be decomposed into five categories. “Within-school:” during year *t-1*, teacher taught in the same school in any assignment other than their assignment in year *t*. “Other school:” during year *t-1*, teacher taught at a different school than their assignment in year *t*. “Previously in district:” teachers who have prior experience in the district but were not present in year *t-1*. “First-year teacher:” teachers whose first appearance in the data occurs during their first year of teaching. “Experienced, other district:” teachers whose first appearance in the data occurs during a year after their first year of teaching. Includes years 2018 to 2025; prior year assignments are undetermined for 2017.

Appendix Figure 1. Event Study Figures From Canonical TWFE Model, Attendance



Note: Each dot represents a separate coefficient on a dummy variable for years since implementation. All non-implementers have a 0 for each dummy variable. Vertical lines represent 95% confidence intervals. Y-axis is relative attendance rate (0-1). Estimates come from the canonical TWFE model with school-by-grade and year fixed effects and no controls. Standard errors are clustered at the school-by-grade level.

Appendix Figure 2. Event Study Figures From BJS Estimator, Attendance



Note: Each dot represents a separate coefficient on a dummy variable for years since implementation. All non-implementers have a 0 for each dummy variable. Vertical lines represent 95% confidence intervals. Y-axis is relative attendance rate (0-1). Estimates come from the BJS estimator with school-by-grade and year fixed effects. Standard errors are clustered at the school-by-grade level.

Appendix A: Student Fixed Effects

Our preferred models will yield causal estimates of NEW so long as there are no within-school-grade factors correlated with the NEW treatment and student achievement. One potential concern is selection at the *student* level: teachers or parents could, for instance, sort students into or out of NEW classrooms based on unobserved factors that influence their achievement growth. For instance, teachers may think that some students would be particularly well-suited to receive team teaching; similarly, parents may try to opt their students into or out of NEW based on their assessments of their children's educational needs. To probe this, we estimate specifications that include student fixed effects, as follows:

$$A_{isgt} = \beta Teamed_{it} + \omega_i + \delta_t + \mu_s + \varsigma_g + \varepsilon_{isgt} \quad (4)$$

where ω_i is a student fixed effect. These within-student estimates compare a student's performance when assigned to a teamed classroom to that same student's performance in years when assigned to a non-teamed classroom. The identifying variation therefore comes entirely from students who *switch* teaming status across years; students observed only in teamed or non-teamed classrooms will not contribute to our estimated effect. This is the key distinction between our preferred specifications, which are identified off the staggered movement of entire school-grades into treatment.

We note two key limitations of the student fixed effects model. First, there is some evidence that these models perform poorly. In particular, Kane et al. (2013) use experimental evidence to demonstrate that the inclusion of student fixed effects can sometimes lead to significant bias in the estimates of teacher effects. Second, the student-switching restriction has particularly large implications for precision in middle school, where we have our most convincingly identified impact estimates. Because middle schools in our district span only two grades (7 and 8), only about half of students are observed in both teamed and non-teamed classrooms in these subsamples, and the students that contribute to the estimated effect are noticeably different on observables.[13] The standard errors rise with the smaller sample, limiting our precision. We report student fixed effects estimates for comparison and transparency, but continue to regard the BJS specification as the more credible impact estimate.

Appendix Table A.1 presents these results. Across grades and subjects, estimates are uniformly small and statistically insignificant. The middle school math estimate is nearly zero (0.004 SD), but the confidence interval is wide enough to contain both zero and estimates from our preferred estimators (roughly 0.08 SD). We interpret this cell with caution given the concerns outlined above: the within-student design, applied to a two-grade window, is too underpowered to distinguish a null from our main effect.

[13] 45% of the individual students in our middle school sample that are ever in a treated classroom do not have an observation in an untreated classroom, and therefore do not contribute to the coefficient in the student fixed effect model. Compared to the other 55% of middle school students that are ever treated, the 45% that are dropped from the estimation have significantly lower standardized test scores (0.05 SD, $p = .002$), significantly lower attendance rates (0.95 percentage points, $p < .001$), and are significantly less likely to be white (3 percentage points, $p = .001$).

Appendix Table A.1. Alternate Specifications Using Student Fixed Effects

	Math			ELA			Attendance		
	Overall	Elem	Middle	Overall	Elem	Middle	Overall	Elem	Middle
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Teamed effect	0.009	0.003	0.004	0.001	0.012	-0.025	-0.000	0.001	-0.004
	(0.018)	(0.024)	(0.058)	(0.012)	(0.017)	(0.030)	(0.001)	(0.001)	(0.003)
N observations	*162264*	*106739*	*33499*	*170668*	*108243*	*41567*	*214611*	*133955*	*65039*
N school-grades	*264*	*229*	*32*	*267*	*229*	*37*	*270*	*230*	*40*

Note: Each cell represents output from a separate regression model using data at the student-year-level. The outcome is identified in each column header. Achievement is standardized within year-grade-subject cells using the district-wide sample. In columns 7-9, the outcome is the proportion of days a student attended in a given year (0-1). The independent variable is a teamed indicator that uses the inclusive definition of teaming within a given year and is equal to 1 for all students in a school-grade-year cell where there is at least one teamed teacher. All regressions include student, school, grade, and year fixed effects. Standard errors are clustered at the school-by-grade level and presented in parentheses. Overall = grades 3-8, Elem = 3-6, Middle = 7-8.
+p < .1; *p < .05; **p < .01; ***p < .001

Appendix B: Sample Restrictions

Our preferred estimates suggest consistent, meaningful impacts of NEW in middle school math, and inconsistent or null impacts on achievement in other grades and subjects. Below, we test whether our preferred estimates are robust to alternate sample restrictions. We focus on two specific data decisions that may warrant scrutiny: the structural change in how treatment status was recorded in the 2025 academic year and the ability of treated units to stop treatment. Appendix Table B.1 presents alternate models for the achievement analysis, testing the robustness of our estimates across more restricted samples in our preferred BJS estimator.

Row 1 limits the sample to observations before 2025, where the teaming list was generated in a meaningfully different way and has markedly lower participation rates. Estimated effects are quite similar to our preferred specification, though the estimates on middle school math are no longer significant given the more limited sample. Interestingly, estimated impacts on ELA are positive and borderline significant ($p = .066$), suggesting there may be some cross-subject middle school effect in the earlier years of our data. More generally, the similarity of effects for middle school math across samples suggests our main estimates are not driven by the different teaming list in 2025.

Row 2 removes school-grade units that moved from treated to untreated at any point in our panel. If anything, the estimated impact on middle school math in this more restricted sample is larger than our preferred specification, though only marginally statistically significant ($p = .054$). Estimates in other grades and subjects are also more muted and closer to zero. Again, units that turn off do not seem to be disproportionately driving our middle school math finding.

Finally, row 3 combines the two restrictions: we limit the sample to observations before 2025 that never turn off. The middle school math effect is no longer significant, which is unsurprising given the much smaller sample, but coefficients are still broadly on par with the estimates from our main model. Again, we see some evidence of a borderline positive middle school effect in ELA (0.058 SD, $p = .092$). Overall, these checks suggest that our preferred estimates are not disproportionately driven by decisions to include observations from 2025 and grades that stop teaming.

Appendix Table B.1. Robustness Checks on Main Achievement Models Using Sample Restrictions

	Math Overall	Math Elem	Math Middle	ELA Overall	ELA Elem	ELA Middle
	(1)	(2)	(3)	(4)	(5)	(6)
Dropping 2025	-0.033	-0.081*	0.083	-0.011	-0.049*	0.059+
	(0.033)	(0.034)	(0.056)	(0.019)	(0.022)	(0.032)
N observations	*165089*	*112844*	*52245*	*171988*	*114008*	*57980*
N school-grades	*260*	*229*	*31*	*264*	*229*	*35*
Dropping grades that stop teaming	0.037	-0.030	0.116+	-0.003	-0.031	0.014
	(0.043)	(0.040)	(0.060)	(0.017)	(0.024)	(0.031)
N observations	*135752*	*90056*	*45696*	*143558*	*91401*	*52157*
N school-grades	*182*	*157*	*25*	*186*	*157*	*29*
Dropping both 2025 & grades that stop teaming	0.016	-0.064	0.086	0.019	-0.059*	0.058+
	(0.054)	(0.049)	(0.070)	(0.020)	(0.025)	(0.035)
N observations	*119592*	*79300*	*40292*	*126502*	*80449*	*46053*
N school-grades	*181*	*157*	*24*	*185*	*157*	*28*

Note: Each cell represents output from a separate regression using data at the student-year level. The outcome is identified in each column header. All cells use the BJS estimator using the did_imputation command in Stata. Row 1 limits the sample to observations before 2025. Row 2 removes grades that ever stopped treatment. Row 3 combines these restrictions. Achievement is standardized within year-grade-subject cells using the district-wide sample. The independent variable is a teamed indicator that uses the inclusive definition of teaming within a given year and is equal to 1 for all students in a school-grade-subject-year cell where there is at least one teamed teacher. All regressions include school-by-grade and year fixed effects. Standard errors are clustered at the school-by-grade level and presented in parentheses. Overall = grades 3-8, Elem = 3-6, Middle = 7-8.
+p < .1; *p < .05; **p < .01; ***p < .001