Extracting Value from Coworkers:

Who Should Work with Whom and When?

John Fallon*

Annenberg Institute, Brown University

john_fallon@brown.edu

August 5, 2026

Abstract

Workers learn from their peers, but which collaborations build lasting skills? I study co-teaching, where general and special education teachers share a classroom, using administrative records on Indiana teachers, 2012–2019. I compare the same teacher before and after co-teaching. Those paired with an experienced partner (16+ years) score 0.10σ higher after returning to solo instruction; estimates for inexperienced partners are near zero. Longer collaborations add nothing. Teachers with two years of experience gain most. Schools do not exploit this: conditional on staffing, partner assignment is indistinguishable from random. Reallocating partnerships toward experienced mentors implies fiscal benefit-cost ratios of 3.6:1–7.0:1.

Keywords: peer effects; coworker spillovers; on-the-job learning; human capital accumulation; worker assignment; teacher productivity; co-teaching; personnel economics

JEL codes: J24, M53, I21, M54, J45

*I declare that I have no relevant or material financial interests that relate to the research described in this paper. This study uses de-identified administrative records provided by the Indiana Department of Education, involves no interaction or intervention with human subjects, and provides no access to identifiable personal information; it is therefore exempt from IRB review.

I thank Andrew Bacher-Hicks, Kevin Lang, Josh Goodman, Marcus Winters, James Feigenbaum, Olivia Chi, Seth Gershenson, Roddy Theobald, Meagan Comb, Sidrah Baloch, Lindsay Kaler, Josefina Senese, Tashnuva Shaheen, Michael Briskin, Danielle Williamson, and Jesse Meredith for helpful comments and suggestions. I also thank participants in the PREREQ seminar. Any remaining errors are my own.

1 Introduction

Peer effects in the workplace are well documented, but how should organizations leverage them to maximize workforce quality? I study this question using co-teaching partnerships, where a general education teacher and a special education teacher share instructional responsibility for a single classroom. This setting allows me to observe the same workers' productivity before, during, and after collaboration. Co-teaching assignment depends on special education enrollment and scheduling constraints rather than teacher quality, providing variation in who collaborates with whom (Section 3). Teachers paired with highly experienced partners gain lasting skills; estimates for teachers paired with inexperienced colleagues are near zero. Longer collaborations do not amplify these gains and may even reduce them. The title phrase "Extracting Value from Coworkers" is intentionally provocative, reframing workplace collaboration in starkly transactional terms. That discomfort is worth it. Better teachers mean better-educated students. Schools already pay for co-teaching. They just aren't claiming this return. Administrators appear to pair teachers almost randomly, and they let partnerships run on long after any benefit has faded.

Existing work shows that coworkers affect each other's output, but it is unclear how that effect is amplified or moderated. One strand finds skill transfer: teachers improve when surrounded by better colleagues (Jackson and Bruegmann, 2009; Nix, 2020; Gershenson et al., 2023). Another points to social pressure or monitoring instead: cashiers work faster when fast peers can watch them (Mas and Moretti, 2009), and workers show up more when their colleagues do (Ichino and Maggi, 2000). Spillovers turn up in other settings too: call-center agents and the coauthors of academic superstars (de Grip and Sauermann, 2012; Azoulay et al., 2010). They are not universal, though: randomly assigned golf partners produce no spillovers at all (Guryan et al., 2009). Proximity alone is not enough; something about the interaction itself has to matter. None of this work asks how long a collaboration should last, or when in a career it pays off most. Those are the questions this paper answers.

Co-teaching is a useful setting for studying how collaboration should be designed. A

general education teacher handles content, and a special education teacher supports a classroom with both disabled and non-disabled students.[1] Schools use this arrangement to satisfy two federal mandates at once: IDEA requires instruction in the least restrictive environment, and NCLB requires a content expert in the room, without either teacher needing dual certification. The two teachers share a classroom every day, which should put more pressure on each other's practice than the occasional interaction typical peer-effects studies rely on. Three things make this setting good for identification: I can observe a teacher's students before, during, and after the partnership; assignment often depends on institutional factors rather than teacher quality; and partnerships dissolve often (over 60% split up within a year) for institutional reasons rather than because the match failed, which gives me variation in collaboration length to work with. A large literature studies how co-teaching affects the students in these classrooms (Austin, 2001; Jones and Winters, 2024; Jones et al., 2025; Bessette, 2008; Cook et al., 2017; De Backer et al., 2023; Andersen et al., 2020; Conderman and Hedin, 2014); I ask how it affects the teachers instead.

Partner experience does most of the work. Teachers paired with a highly experienced colleague (16+ years) show 0.10σ higher student achievement after returning to solo instruction, comparable to the effect of an intensive coaching program (Kraft et al., 2018). Teachers paired with an inexperienced colleague (0-7 years) show estimates near zero. This pattern is consistent with experienced partners holding knowledge that transfers through daily contact.

Longer partnerships do not produce larger gains. Comparing teachers who co-taught for one, two, or three years, I find no evidence that extra time together helps. If anything, extended partnerships may hurt, by coming at the expense of the solo teaching experience that otherwise builds a teacher's skill.

Career timing also matters. Teachers with two to three years of prior solo experience benefit most; teachers with only one year gain less. Some baseline competence may be needed before a teacher can learn from an experienced partner.

[1] Models of how teachers split responsibility vary. I am agnostic about what model the teachers follow.

Experience itself improves productivity, though why remains unclear (Arrow, 1962; Chen, 2021; Haggag et al., 2017; Ost, 2014; Shaw and Lazear, 2008; Thompson, 2010, 2012). My results suggest collaboration can speed this process up, but only under the right conditions. The closest work is Papay et al. (2020), who show experimentally that pairing teachers by specific skill dimensions raises performance. I add two margins their design doesn't address: how long a partnership should last, and when in a career it pays off most.

To translate these estimates into policy terms, I simulate strategic pairing against current practice, tracking teachers over their careers and accounting for the cost of pulling an experienced mentor out of their own classroom. Under baseline assumptions, strategic pairing generates the equivalent of $4,021 in additional lifetime earnings per student in a school, with positive effects emerging within 6 years. Even under the most pessimistic assumptions I consider, cumulative effects turn positive within 8 years, and the opportunity cost would need to be twice its empirically estimated size before the policy stopped paying off.

Scaled across a workforce, strategic pairing could raise student achievement at low cost, since it primarily reallocates existing staff. Yet schools rarely do this: assignment tracks scheduling convenience, not developmental value, and novice-novice pairings stay common even though they produce no gains. Firms in this position should treat pairing as a deliberate choice, not a scheduling byproduct: put experienced workers with junior colleagues, briefly, and time it well.

The same tradeoffs likely apply outside teaching: which pairings help, how long they should last, and when in a career they matter most. Hospitals assigning residents to attending physicians, consulting firms staffing project teams, and tech companies structuring code review all face this choice, usually by default rather than by design. The results here suggest that the *structure* of a collaboration, not just its existence, is what determines whether it builds skill. Whether these specific numbers hold in other industries is an open question, but the underlying logic, that pairing choices are a design problem worth getting right, should travel.

2 Data

I use longitudinal administrative data from the Indiana Department of Education covering all students and teachers from 2012 to 2019. I restrict the sample to traditional public schools, excluding charter and private schools. A unique classroom identifier links each student to every teacher assigned to their classroom. This is what allows me to identify classrooms with more than one teacher. Student records include annual demographics, supplemental service classifications (special education, English language learners), and standardized test scores in math and ELA. I standardize scores by grade and year. Teacher records include age, experience, degree, and licensure area. They also include an indicator for a teacher's first year in Indiana and a role designation that flags special education teachers as team teachers or co-teachers. Indiana tests students annually in grades 3 through 8. Because the analysis controls for two years of prior scores, the estimation samples consist of students in grades 5 through 8.

I define a co-taught classroom as a class with exactly two assigned teachers, one of whom is flagged as a team teacher or co-teacher. I treat the flagged teacher as the special education teacher and the unflagged teacher as the general education teacher. The flag is what distinguishes genuine pairings from mid-year replacements. Without it, a classroom in which one teacher left and another took over would look the same as a co-taught classroom in the data. I drop classrooms with three or more assigned teachers. In those classrooms I cannot separate replacements from true pairings, and a large teacher count often signals a mislabeled delivery model such as a pullout classroom.

I study two samples, both summarized in Table 1. The *co-taught sample* is every co-taught classroom and the teachers and students in it. The *event study sample* is the sample that I use to measure skill development, and it has two arms. The treated arm is a subset of the co-taught sample: general education teachers whom I observe both before their first co-teaching year and after their return to solo instruction. The control arm is drawn from teachers who never co-teach and who match the treated teachers on career profile. The

event study sample is more selected than the co-taught sample, so I describe below how the two differ. Appendix Table A.2 traces the restrictions that reduce the teacher panel to this sample.

Measuring development requires each teacher's true experience, so I keep only teachers whom I observe from their first year in the data, identified by Indiana's "First Year" flag. The administrative "Years of Experience" field is not a substitute. It ignores career breaks, so it misclassifies teachers who return after time away. The flag is available for 5,943 of the 29,630 teachers in the panel, so the restriction drops 80 percent of the workforce (Table 1). The teachers I drop are excluded because of when they entered, not because of anything about their quality. For a teacher already in a classroom in 2012, the career began before the panel does, so experience is left-censored and I cannot measure it.

The excluded teachers differ from the rest of the state on the margins this cohort restriction predicts, and are otherwise similar. The no-known-first-year group is more experienced (13.2 versus 10.8 years), older (43.5 versus 41.0), and more credentialed (52 versus 44 percent hold a master's). They are demographically similar to the full workforce, at 79 percent female and 92 percent white in both groups (Appendix Table A.1). The teachers who remain are therefore younger and less experienced than the average Indiana teacher. This bounds the external validity of the average effect to recent-entry, early-career teachers, which is the population the development analysis restricts to in any case (Section 6). The headline experienced-partner gradient is not bounded in the same way. I relax the restriction in a robustness check that imputes entry for the excluded veteran cohorts, and the gradient persists among teachers who co-teach at any career stage (Figure A.18).

Within this population, co-teaching is common: 24 percent of teachers co-teach at some point, including 20 percent of general education teachers and 54 percent of special education teachers. Teachers who ever co-teach look similar to those who never do on experience at entry (Appendix Table A.1). The development analysis follows general education teachers, so the partners whose experience defines the headline gradient are special education teachers. The estimates therefore measure transfer from experienced special educators to early-career

general educators. The event study sample also looks different on the student side. Students in it are more advantaged on average than students in the full co-taught sample, because less-experienced teachers tend to be assigned higher-achieving, less disadvantaged classrooms (Table 1).

3 Co-Teaching Assignment Patterns

This section establishes two facts that the identification strategy depends on. First, co-teaching assignments are driven by institutional constraints, not by judgments about teacher quality. This applies both to who co-teaches and to who they are paired with. By "institutional constraints" I mean the mechanical forces that govern staffing: special education enrollment, scheduling, and certification rules. Second, student composition shifts predictably when a teacher begins co-teaching, which is why the later analysis controls for prior achievement.

3.1 Entry and Exit Patterns Reflect Institutional Constraints

Both special education and general education teachers are most likely to begin co-teaching in their first few years, and the hazard declines with experience (Figure 1). This pattern is consistent with junior teachers having less seniority-based control over their classroom assignments.

Co-teaching sections are created in response to special-education enrollment. Within a school-grade cell, an increase in the number of special-education students raises the probability that the grade is co-taught (Appendix Figure A.1). The response is concentrated in grades that start with few special-education students and fades in grades that already have many. Sections are added as enrollment need arises.

Partnerships dissolve quickly. Over 60% of pairs separate after one year together, and the separation probability remains above 40% after five years (Figure 2). The hazard falls with tenure but stays substantial. Most partnerships are therefore short, and most general

Table 1: Summary Statistics

Panel A: Sample Size

	Co-Taught Sample	Event Study Sample
Number of Teachers	10,100	804
Treated (ever co-taught)		141
Control (never co-taught)		663
Number of Students	152,847	61,652
Number of Classes	24,129	7,361

Panel B: Teacher Characteristics

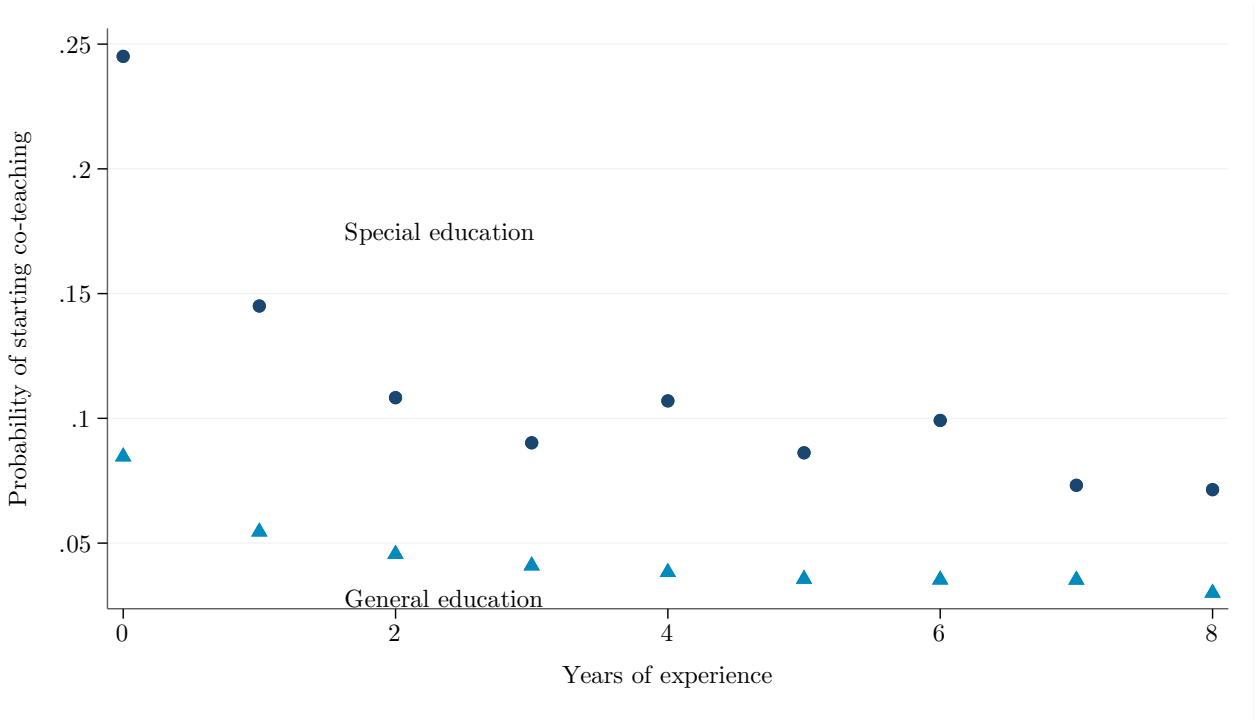
	Never Co-Taught	Ever Co-Taught	No Known First Year	All Indiana
Number of Teachers	3,425	1,175	23,687	29,630
Years of Experience (Mean)	1.15	1.38	13.20	10.85
General Education Share	0.91	0.72	0.87	0.86
Ever Switches Gen-Ed/Sp-Ed Role	0.00	0.06	0.02	0.02
Master's Degree or Higher	0.12	0.12	0.52	0.44
Age (Mean)	31.2	31.2	43.5	41.0
Female	0.75	0.77	0.79	0.79
White	0.90	0.90	0.92	0.92
Black	0.06	0.06	0.06	0.06
Hispanic	0.02	0.02	0.01	0.01

Panel C: Student Characteristics

	Co-Taught Sample		Event Study Sample	
	Mean	SD	Mean	SD
Test Z-Score	−0.387	0.965	−0.141	0.986
Special Education	0.329		0.227	
English Language Learner	0.097		0.089	
Free/Reduced-Price Lunch	0.606		0.506	
Female	0.474		0.487	
White	0.600		0.688	
Black	0.193		0.124	
Hispanic	0.133		0.116	
Grade (Mean)	5.0		6.2	
Observations (Test Scores)	455,759		171,843	
Observations (Demographics)	630,310		171,843	

Notes: Indiana administrative data, 2012–2019. *Panel A*: the co-taught sample is all teachers and students in classrooms with exactly two assigned teachers (one general, one special education). The event study sample contains both experimental arms of the stacked design: the treated teachers are the general education teachers meeting the development-analysis criteria (known first year, 1–3 years of prior solo experience, an observed post-co-teaching solo year), and the controls are never-co-taught teachers who match a treated cohort exactly on entry year, exit year, and subject. Both arms are subsets of the corresponding Panel B populations (141 of the 1,175 ever-co-taught; 663 of the 3,425 never-co-taught): the event study additionally requires classrooms in tested grades and subjects, an exactly matching career window, and balanced observation within each event. *Panel B* reports teacher-level means; the never/ever columns are the development analysis sample, the no-known-first-year and all-Indiana columns the full teacher panel, so the four are not additive. Appendix Table A.1 adds standard deviations and difference tests. *Panel C* reports student-year means; observations are listed separately for test scores and demographics.

Figure 1: Hazard Rate for Entering Co-Teaching



This figure plots the annual probability that a teacher begins co-teaching for the first time, conditional on years of experience and not having co-taught previously. The sample includes all Indiana teachers from 2012–2019 identified with the "First Year" flag. Rates are computed separately for special education and general education teachers. The declining hazard is consistent with junior teachers having less seniority-based control over classroom assignments.

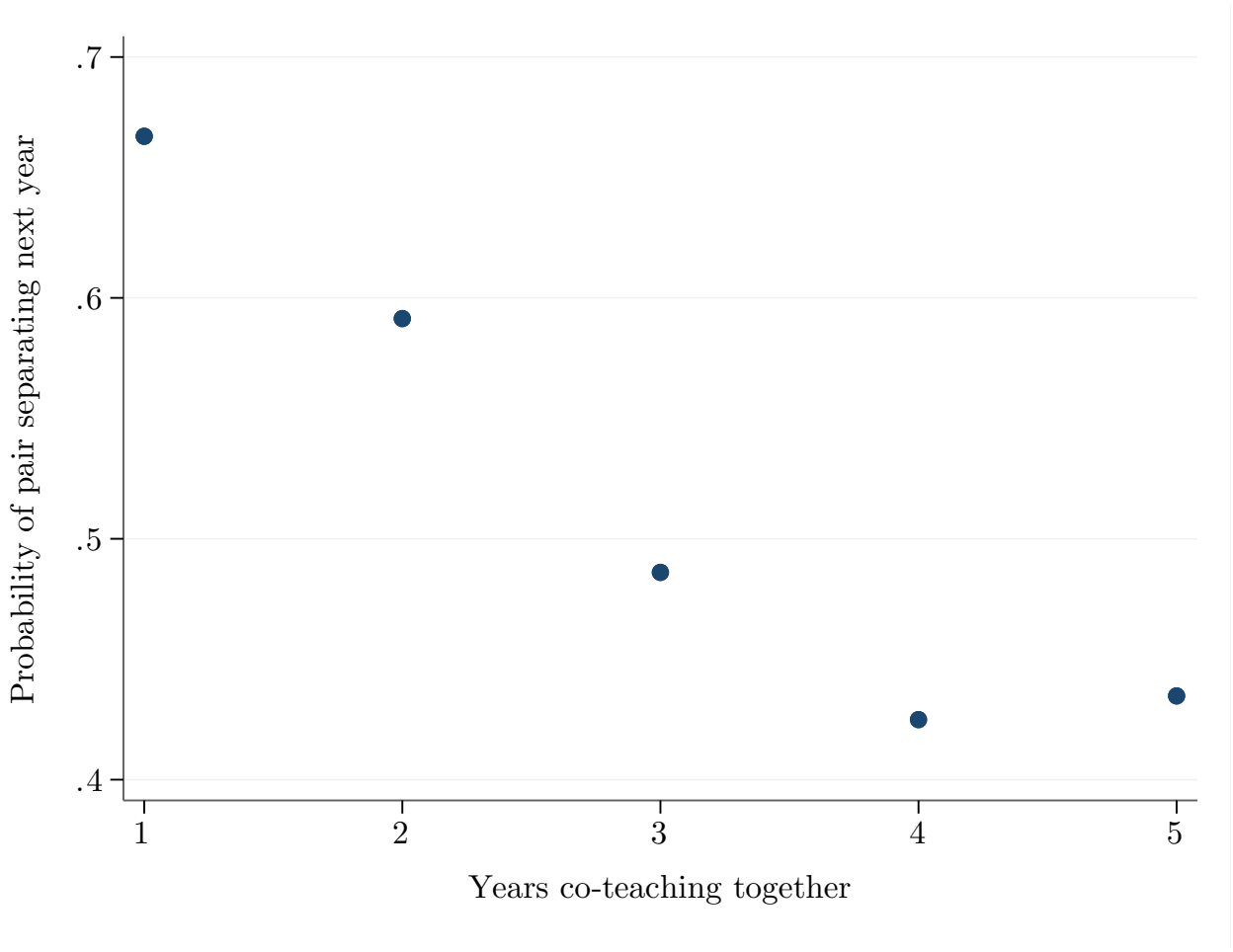
education teachers who co-teach return to solo instruction (Appendix Figure A.4), which is what the event study design requires.

General education and special education teachers follow different paths after partnerships end (Figures A.4 and A.5). General education teachers typically return to solo instruction (76%), while special education teachers cycle to new co-teaching partners (47%). The accumulation of partners reflects the same asymmetry: the average co-teaching special education teacher works with 2.9 distinct partners over the sample period, while the average general education co-teacher works with 1.3.

3.2 Student Assignment Changes Predictably

Teachers who co-teach receive different students than solo teachers. I test this using difference-in-differences comparing student assignment before and after co-teaching begins

Figure 2: Hazard Rate for Co-teaching Pair Separation



This figure plots the annual probability that a co-teaching pair separates in the following year, conditional on the number of consecutive years together. The sample includes all co-teaching pairs observed in Indiana from 2012–2019. Separation is defined as the two teachers no longer sharing a classroom in the subsequent year. The vertical axis spans the observed range of the hazard rather than starting at zero.

(Table 2). Here "post" means any year from a teacher's first co-teaching year onward. Because 76% of general education teachers return to solo instruction within a year, post years mix co-taught and solo classrooms. This makes the test conservative, understating the difference between co-taught and solo classrooms.

General education teachers who co-teach receive 2.0 percentage points more students with disabilities and 2.4 percentage points more students eligible for free or reduced-price lunch after beginning co-teaching. The first shift follows from the purpose of a co-taught classroom, which exists to serve students with disabilities; the second follows from the association between disability status and economic disadvantage. English Language Learner assignment

does not change significantly.

The pooled "post" effect combines the years a general education teacher is actively co-teaching with the years after returning to solo instruction. Panel B of Table 2 decomposes the interaction into these two states, and the composition shift is entirely contemporaneous. While actively co-teaching, classrooms hold 4.0 percentage points more students with disabilities (s.e. 0.2) and 4.6 points more free-lunch-eligible students (s.e. 0.4). Once teachers return to solo instruction, neither difference is distinguishable from zero (special education −0.0pp, s.e. 0.3; free lunch 0.7pp, s.e. 0.5). The post-period achievement estimates of Section 5 are identified largely from these solo-return years. In those years, a teacher's classroom looks on observables like it did before they co-taught. Any persistence in those years therefore reflects the teacher rather than a measurable difference in who they teach.

I control for these observable student characteristics in all analyses. Unobservable sorting is harder to rule out, but the event study design in Section 5 tests for parallel pre-trends, which would detect unobservable changes that begin before co-teaching does.

3.3 Partner Assignment Shows Minimal Strategic Pairing

Administrators do not systematically pair experienced teachers with novices. I compare observed pairing frequencies to what would occur under random assignment (Figures A.6 and 3).

The deviation from random pairing runs *against* strategic mentorship. Novice-novice pairings, where both teachers have at most two years of experience, are 0.8 percentage points *more* common than random within-year pairing of the available teachers would predict, 23 percent above the expected frequency of 3.5 percent. Inexperienced teachers are not being routed to experienced partners. They cluster together, because inexperienced teachers are the most likely to co-teach and the most likely to work in the same schools.

The formal tests find little structure beyond that. A chi-square comparison of observed cell frequencies to the within-year random benchmark does not reject randomness at the five

Table 2: Difference-in-Differences on Student Assignment by Co-Teaching Status

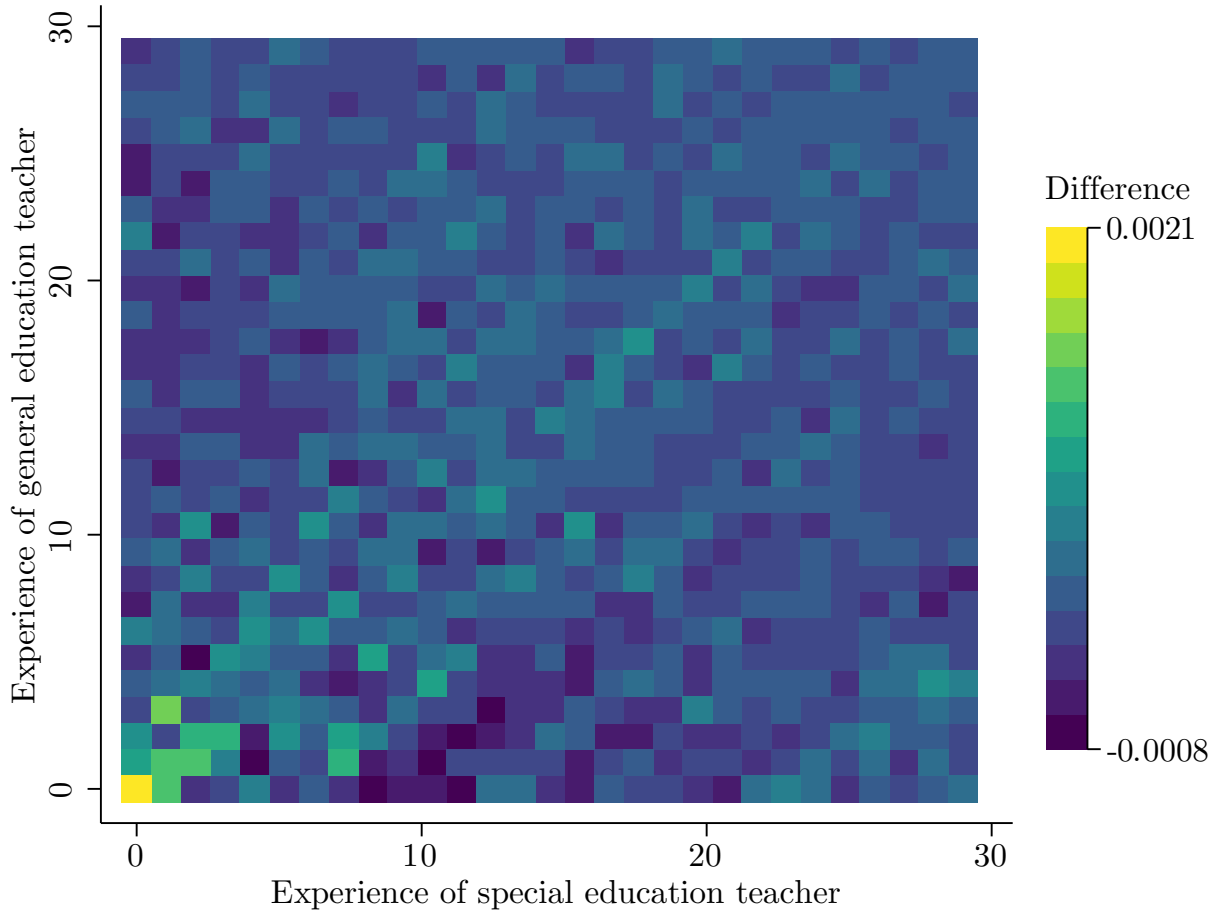
	Share Special Ed			Share FRPL		
	All	Gen Ed	Spec Ed	All	Gen Ed	Spec Ed
Panel A: Pooled Post-Period						
Ever Co-taught	0.008	-0.007^{**}	-0.090^{***}	0.025^{***}	0.022^{***}	0.002
	(0.005)	(0.002)	(0.011)	(0.004)	(0.004)	(0.009)
Ever Co-taught × Post	−0.006	0.020^{***}	-0.312^{***}	0.024^{***}	0.024^{***}	-0.018^{*}
	(0.005)	(0.002)	(0.012)	(0.004)	(0.004)	(0.009)
Panel B: Post-Period Decomposed						
Ever Co-taught × Actively Co-teaching	0.011^{*}	0.040^{***}	-0.355^{***}	0.041^{***}	0.046^{***}	-0.025^{**}
	(0.005)	(0.002)	(0.011)	(0.004)	(0.004)	(0.009)
Ever Co-taught × Returned to Solo	-0.020^{***}	−0.000	-0.096^{***}	0.009	0.007	0.027^{*}
	(0.005)	(0.003)	(0.016)	(0.005)	(0.005)	(0.011)
Observations	210,263	181,736	28,411	210,263	181,736	28,411
Dep. Var. Mean	0.232	0.159	0.695	0.507	0.489	0.622

Notes: This table estimates changes in student composition assigned to teachers before and after they begin co-teaching. In Panel A, the specification regresses classroom-level student shares on an indicator for ever co-teaching, an indicator for the post-co-teaching period, and their interaction, absorbing teacher-experience and school-by-year fixed effects. Panel B re-estimates the same specification splitting the post-period indicator into years the teacher is actively co-teaching and years after returning to solo instruction; the Ever Co-taught main effect (not shown) is essentially unchanged. The sample includes all Indiana teachers from 2012–2019 identified with the "First Year" flag. Columns separate by teacher type (all, general education only, special education only). The dependent variable is the share of students in the teacher's classroom with the indicated characteristic. Standard errors in parentheses are clustered at the teacher level.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

percent level ($\chi^2(878) = 943$, $p = 0.06$, among cells with expected counts of at least five). What deviation there is reflects where novices work, not how administrators pair them. Rebuilding the benchmark with each school's own experience mix as the marginals moves the test further from rejection ($\chi^2(854) = 722$, $p = 1.00$), and conditioning on school-by-year cells fits tighter still ($\chi^2(860) = 352$, $p = 1.00$). The novice-novice excess falls from 0.8 percentage points statewide to 0.4 within schools and 0.0 within school-years. Conditional on a school's available staff, pairing is indistinguishable from random. Identification in Section 5 does not rest on pairing being literally random; it requires that partner experience be unrelated to the focal teacher's counterfactual development, which the pre-trend and composition checks address directly.

Figure 3: Difference From Random Pairing of Teachers



This heatmap displays the difference between observed co-teaching pairing frequencies by experience level and the frequencies expected under random within-year pairing of the available teachers: for each cell, the expected share is the year-weighted product of that year's general education and special education experience distributions, so the benchmark reflects who was actually available to pair in each year. Each cell shows the percentage point difference (observed minus expected) for a given combination of general education teacher experience (x-axis) and special education teacher experience (y-axis). Positive values indicate pairings that occur more often than random; negative values indicate less often. The sample includes all co-teaching pair-years in Indiana from 2012–2019. Figure A.6 shows the raw pairing frequencies.

Identification does not require that pairing be random. It requires that teachers who draw experienced partners were not already on a different achievement path, in levels or in trend. Neither appears in the data. Partner experience is unrelated to a teacher's mean pre-co-teaching student achievement (event time -3 to -1): the coefficient is 0.39 years per standard deviation, with a confidence interval that rules out sorting larger than 2.30 years (Table A.3). An indicator for drawing a partner with 16 or more years of experience gives the same answer. Teachers who go on to pair with highly experienced partners also have flat pre-trends (Figure 6), so the post-period gain is not a pre-existing divergence continuing.

4 Do Pairs Develop Match-Specific Skills?

Co-teaching could build match-specific capital, transferable individual skills, or both. Match-specific capital makes a particular partnership productive but is lost when the partners separate, while transferable skills stay with the teacher. This section tests for match-specific capital; Section 5 takes up individual skill development.

4.1 Identification: Separating Pair Tenure from Experience and Time

Within any partnership, pair tenure, each teacher's total experience, and the calendar year all advance one-for-one. This is an exact collinearity: a regression of student outcomes on pair tenure with pair fixed effects cannot separately identify tenure, experience, and year effects. It is the same problem that arises in estimating returns to teacher experience.

I adapt the two-step solution of Papay and Kraft (2015).[2] The first step estimates calendar-year effects from a specification that excludes pair fixed effects,

$$\pi_{ipy} = \lambda_y + \sum_k \delta_k \, 1[\text{tenure}_{py} = k] + X'_{ipy}\beta_1 + \varepsilon_{ipy}, \tag{1}$$

where π_{ipy} is the test score of student i taught by pair p in year y and X contains the student controls. The tenure indicators enter with the same functional form as in the second step, so that the year effects do not absorb tenure variation. The δ_k are discarded. Year effects λ_y are identified because pairs form in every year of the panel, so each year's cross-section mixes new and established partnerships. The second step subtracts $\hat{\lambda}_y$ from scores and regresses the adjusted outcome on pair-tenure indicators with pair fixed effects,

$$\pi_{ipy} - \hat{\lambda}_y = \sum_k \theta_k \, 1[\text{tenure}_{py} = k] + \mu_p + X'_{ipy}\beta_2 + u_{ipy}, \tag{2}$$

[2] A single-step alternative that imposes the year effects and tenure profile simultaneously produces similar results (Figure A.7).

estimated with and without controls for the two teachers' total experience. Tenure and experience advance one-for-one within a pair, so the pair fixed effects leave no within-pair variation that could separate them. The separation instead comes from a cross-pair restriction: the experience controls are constrained to a single profile common to all pairs. Because pairs form at different points in their members' careers, a given tenure year is observed at many different experience levels across pairs, and the tenure coefficients are identified from that variation. Standard errors are from a pair-clustered bootstrap that re-runs both steps, so they account for estimation of the year effects in the first step.

The approach rests on two assumptions. First, year effects are common across pairs of different tenures. This would fail if aggregate shocks hit established pairs differently than new ones. Second, within-pair student composition does not trend with tenure conditional on the controls. I compare student composition in short- and long-duration pairs in Appendix Table A.4, and the composition controls enter throughout.[3]

4.2 Longer-Lasting Pairs Are Not Better Matches

Longer partnerships do not produce better student outcomes once teacher quality is held fixed. I regress student test scores on a partnership's completed duration, with and without separate fixed effects for each teacher in the pair (Table 3).

Without teacher fixed effects, longer-lasting pairs associate with higher test scores, and the association is strongest in math. Panel B adds separate fixed effects for each member of the pair, so the comparison holds each teacher's quality fixed. Because completed duration does not vary within a partnership, these coefficients come from teachers who appear in more than one partnership. Repeat partnerships are common enough to support the comparison: special education teachers participate in 2.9 partnerships on average, and general education

[3] A third assumption is that the return to experience is common across pairs, since the experience controls are constrained to a single profile. If teachers who partner early in their careers improve at different rates than those who partner later, the tenure coefficients may absorb the difference. Allowing separate experience profiles for early- and late-forming pairs leaves every tenure coefficient within a standard error of its common-profile estimate (Appendix Table A.5).

teachers in 1.3. The association then disappears for English and turns negative in math, where an additional year of partnership duration predicts 0.031 SD lower scores (Table 3). The positive association in Panel A comes from better teachers staying in partnerships longer. Time in a partnership does not make a pair better, and in math it is associated with worse outcomes.

Table 3: Regression of Student Test Scores on Completed Partnership Duration

	Student Test Scores		
	Combined	Math	English
Panel A: Without Teacher Fixed Effects			
Years Spent as Pair (Total)	0.00809*	0.0125**	0.00649
	(0.00365)	(0.00421)	(0.00470)
Observations	142,805	61,646	81,143
Panel B: With Individual Teacher Fixed Effects			
Years Spent as Pair (Total)	−0.00455	−0.0313*	0.000386
	(0.00688)	(0.0121)	(0.0105)
Observations	142,525	61,408	80,903

Notes: This table regresses student test scores (standardized by grade and year) on the total number of years a co-teaching pair worked together (the partnership's completed duration). Panel A includes school, year, and experience fixed effects only. Panel B adds separate individual fixed effects for the general education and special education teacher, rather than a joint pair fixed effect, which would be perfectly collinear with completed duration. Because completed duration does not vary within a partnership, the Panel B coefficient is identified from teachers who appear in multiple partnerships, comparing partnerships of different completed lengths that involve the same teacher (special education teachers average 2.9 partnerships; general education teachers, 1.3). The sample includes all students in co-taught classrooms in Indiana from 2012–2019. Controls include student demographics and a cubic polynomial in the previous two years' test scores. Standard errors in parentheses are clustered at the pair level.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3 Pairs Do Not Improve Together Over Time

Teachers do not learn to work together more effectively over time. I test this using the two-step approach of Section 4.1, which separates pair-specific learning from general experience and year effects.

Without controlling for general teaching experience, student test scores increase with

years together as a pair (Figure 4). The association goes away once experience controls are added. Teachers improve over time, but there is no evidence they improve faster in a pair than teaching solo.

With experience controls the point estimates lean *downward* with pair tenure rather than sitting at zero, though the profile is not statistically distinguishable from flat (joint $p = 0.66$). The pattern is suggestive, and it points the same way as the negative math coefficient in Table 3. Both are consistent with time in a partnership coming at the expense of solo teaching experience, a mechanism I develop in Section 5.

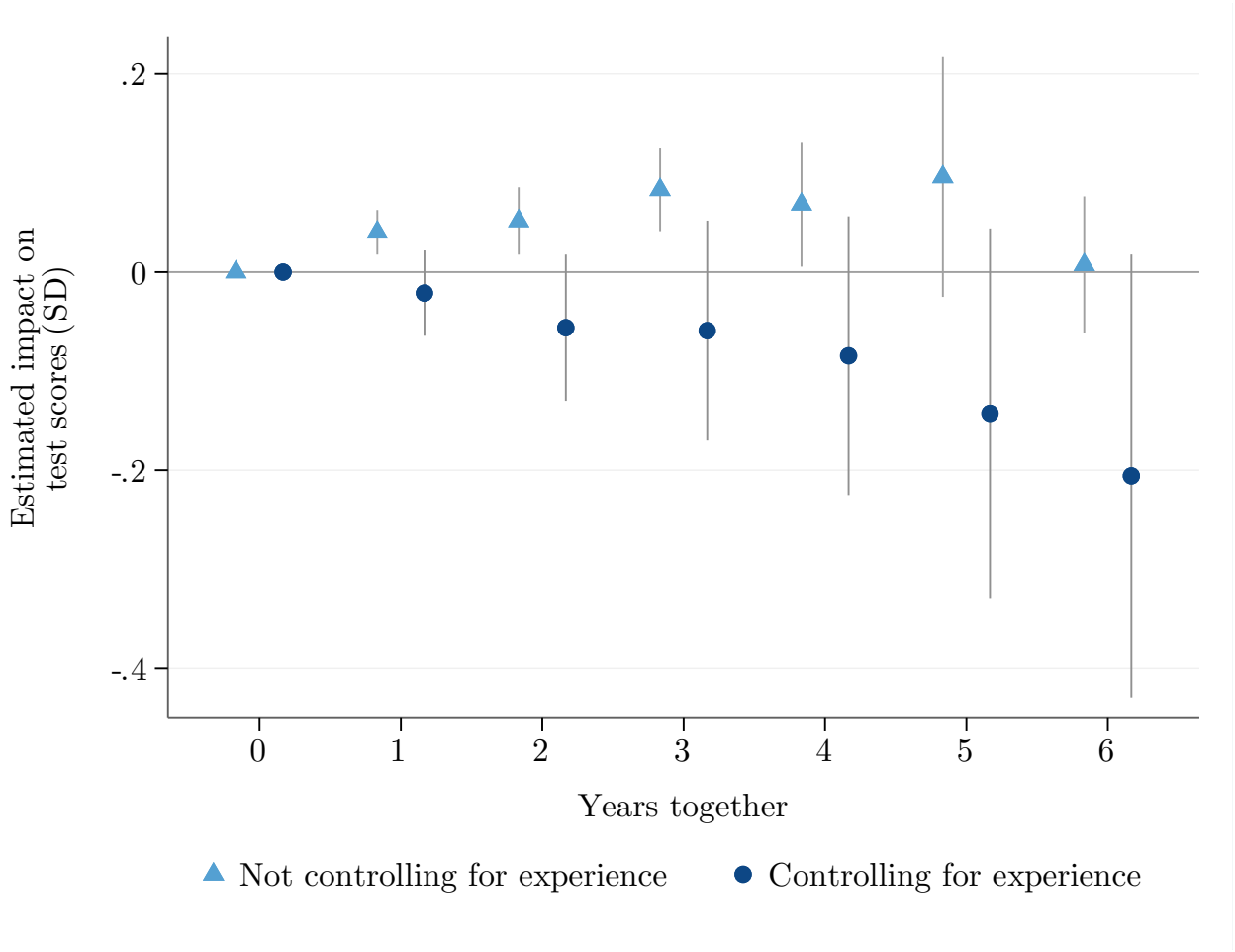
An alternative reading is that pairs who stay together for many years are negatively selected, kept together because one or both teachers struggle solo. The evidence cuts against this reading. The cross-sectional association between duration and scores is *positive* until teacher quality is held fixed (Table 3). The within-pair estimates also condition on pair identity, so selection would have to work through time-varying shocks rather than the fixed quality of the match. Pairs that last three or more years have *more* experienced members, by roughly 2.4 years for the general education teacher and 3.5 for the special education teacher, and their students score slightly higher (Appendix Table A.4). On observables, long-duration pairs are positively selected, not negatively.

There is little evidence that co-teaching builds match-specific capital. The next section tests whether it instead accelerates individual skill development that persists after partners separate.

5 Solo Teaching After Co-Teaching

Section 4 found no evidence that time together makes a partnership more productive. This section asks whether co-teaching instead changes the teacher, leaving skills that persist after the partnership ends. I follow general education teachers before and after their first co-teaching year, comparing them to teachers who never co-taught but share the same career timeline. Their students score higher afterward, including in years the teacher is back to

Figure 4: Return To Experience in a Pair



This figure plots estimated returns to years spent together as a co-teaching pair, using the two-step method from Papay and Kraft (2015). Step one estimates calendar-year effects from a specification with student controls that excludes pair fixed effects. Step two subtracts the estimated year effects from test scores and regresses the adjusted outcome on pair duration indicators with pair fixed effects, with and without controlling for total teaching experience. Controls include student demographics and a cubic polynomial in the previous two years' test scores. The omitted category is the pair's first year together. The sample includes all co-teaching pairs in Indiana from 2012–2019. Bars are 95% confidence intervals from a pair-clustered bootstrap that re-runs both steps.

teaching alone. How large the gain is depends on the partner's experience and on the teacher's own; longer partnerships add nothing, and the estimates suggest they subtract.

5.1 Estimation Strategy

The obstacle is that teachers start co-teaching in different years. Comparing a teacher who just started to one who started years earlier produces biased estimates when the effect builds or fades over time (Goodman-Bacon, 2021). I use the stacked event-study design of Cengiz et al. (2019), which sidesteps those comparisons by building a separate, self-contained

experiment around each treated cohort and then stacking the experiments together. A remaining concern is non-convex weighting (Wing et al., 2024). Section 6 addresses it two ways. I show the estimate is stable across explicit weighting schemes. I also re-estimate the design with the imputation estimator of Borusyak et al. (2024), which is free of non-convex weighting by construction and accommodates the prior-score controls this setting needs.

Each "event" is one such experiment. I first group treated teachers who share four traits: the subject they teach, the year they started teaching, the year they started co-teaching, and the year they left teaching. I call this set of traits a matching profile. For each treated group I then assemble a control group of teachers who *never* co-taught but match on subject, entry year, and exit year. I place these controls on their cohort's event-time clock, where time zero is the cohort's first co-teaching year, so that "year -1" and "year $+1$" denote the same career distance for treated and control teachers alike. They serve only to show what test scores would have done, absent co-teaching, for a cohort on the same career timeline. One never-co-taught teacher can match several treated cohorts, so the same teacher and the same student-year can appear in more than one event. I therefore cluster all standard errors at the teacher level.

One event consists of math teachers who entered teaching in 2014, first co-taught in 2016, and remained through 2019. Its controls are math teachers who entered in 2014, remained through 2019, and never co-taught. Event time zero is 2016 for everyone in the event, so year -1 is 2015 and year $+1$ is 2017 for both groups. I estimate the treatment effect by regression, comparing how the treated teachers' student outcomes move relative to the controls' around 2016, with fixed effects for each teacher and for each event-year.

I restrict treated teachers to those with one to three years of experience before co-teaching. Early entry is the institutional default. Teachers who first co-teach after four or more years show declining scores before co-teaching begins (Figure A.8), which suggests their assignment responds to performance. Section 6 develops the argument. Including all teachers with one or more years of prior experience preserves the pattern (Figure A.17).

On the stacked data I estimate:

$$\pi_{ditsy} = \sum_{\substack{k=-3 \\ k\neq -1}}^{2} \omega_k \, YTC_{tyk} + X'_{isy}\alpha + \delta_{dt} + \gamma_{dy} + \Phi_{isy} + \epsilon_{ditsy} \tag{3}$$

where π_{ditsy} is student i's test score in event d for teacher t in subject s in year y, and YTC_{tyk} indicates that year y is k years from teacher t's first co-teaching year, with the year before entry ($k = -1$) omitted as the reference period.

The remaining terms are the control set. I describe what each one does, because several of the exercises below turn on what they absorb.

The vector X_{isy} does two jobs. Its first part is a cubic polynomial in each of the student's previous two years of math and ELA scores. This is what strips out the part of measured achievement that comes from which students a teacher was assigned rather than from how well the teacher taught them. Each term is interacted with the subject being tested, because a prior math score predicts a math outcome differently than it predicts an ELA outcome, and each carries an indicator for a missing score so that students with incomplete testing histories stay in the sample rather than dropping out of it. The second part is the composition of the classroom itself: the shares of students who are special education, English learners, eligible for free or reduced-price lunch, Black, and male. An individual student's own score history says nothing about the difficulty of the room they sit in, and co-teaching changes that room (Section 3).

The two event-specific fixed effects define what is being compared to what. The term δ_{dt} gives each teacher their own baseline within each sub-experiment, so the coefficients trace how a teacher's students change relative to that same teacher's earlier students, not relative to teachers who were better or worse to begin with. The term γ_{dy} absorbs whatever moved all classrooms in a given year, such as a change in the test instrument or in state policy.

The term Φ_{isy} collects the remaining sources of variation that could move test scores for reasons that have nothing to do with the teacher. Class size enters on its own, because a

teacher whose classes happen to shrink after co-teaching would post higher scores without having become any better at the job. Grade-by-subject effects handle the fact that tests differ in difficulty and in spread across grades and subjects even after I standardize them. Student ethnicity and gender, along with indicators for whether a student was ever classified as special education or as an English learner, absorb persistent differences between students that two years of scores do not fully capture. Status-transition effects go a step further: they separate a student who has just been classified into special education, or who has just lost free-lunch eligibility, from a student whose status has been stable for years, since the two are on different trajectories even when their current classification is identical. Finally, a school-by-free-lunch-status effect gives each school its own achievement gap between economically disadvantaged students and everyone else, so that a teacher is neither credited nor penalized for working in a building where that gap happens to be unusually wide.

Every specification below uses this same control set unless I say otherwise.

Outcomes come only from classrooms the focal teacher teaches alone. Their co-taught section is excluded, so even in year zero the estimate reflects performance in a classroom with no second teacher.

One caution about interpretation. Event time is measured from a teacher's *first* co-teaching year, so the post-period coefficients capture outcomes after co-teaching *began*, not after it ended. In year zero the teacher is still in a partnership, even though the classrooms I measure are their own. In later years some have returned to solo instruction while others continue or start new partnerships. I therefore read the post-period estimates as the effect of having co-taught, averaged over both states. I verify separately that the gains remain once I condition on whether a teacher is currently co-teaching (Figure A.9).

5.2 Teachers Improve After Co-Teaching

Teachers show higher student achievement after co-teaching than before (Figure 5). Before co-teaching, treated and control teachers follow parallel trends, with small and

statistically insignificant differences in years −3 through −1. If co-teachers had been on steeper improvement trajectories, they would have diverged from controls before co-teaching began. The gain appears with the onset of co-teaching. In the first co-teaching year, students score 0.04σ higher than the matched controls'. In the two following years, once teachers have largely returned to solo instruction, the estimates are 0.01σ and 0.07σ.[4] A joint test does not reject that the three are equal ($p = 0.17$), so the year-to-year movement is sampling variation around a common effect of roughly 0.04σ rather than a dip and rebound. Year zero is a year of active co-teaching, but the classrooms I measure have no second teacher in them, so the year-zero estimate is not a contemporaneous staffing effect.

The composition of the post-period years points the same way. For teachers whose co-teaching lasted exactly one year, the partnership ends after year zero, so years 1 and 2 contain no active partnership. Their estimates are 0.043σ (s.e. 0.026), 0.051σ (0.034), and 0.078σ (0.055) in years 0 through 2, in line with the pooled estimates though individually imprecise. Teachers with 2+ year spells show smaller and noisier effects, consistent with the finding that additional years of co-teaching add nothing.

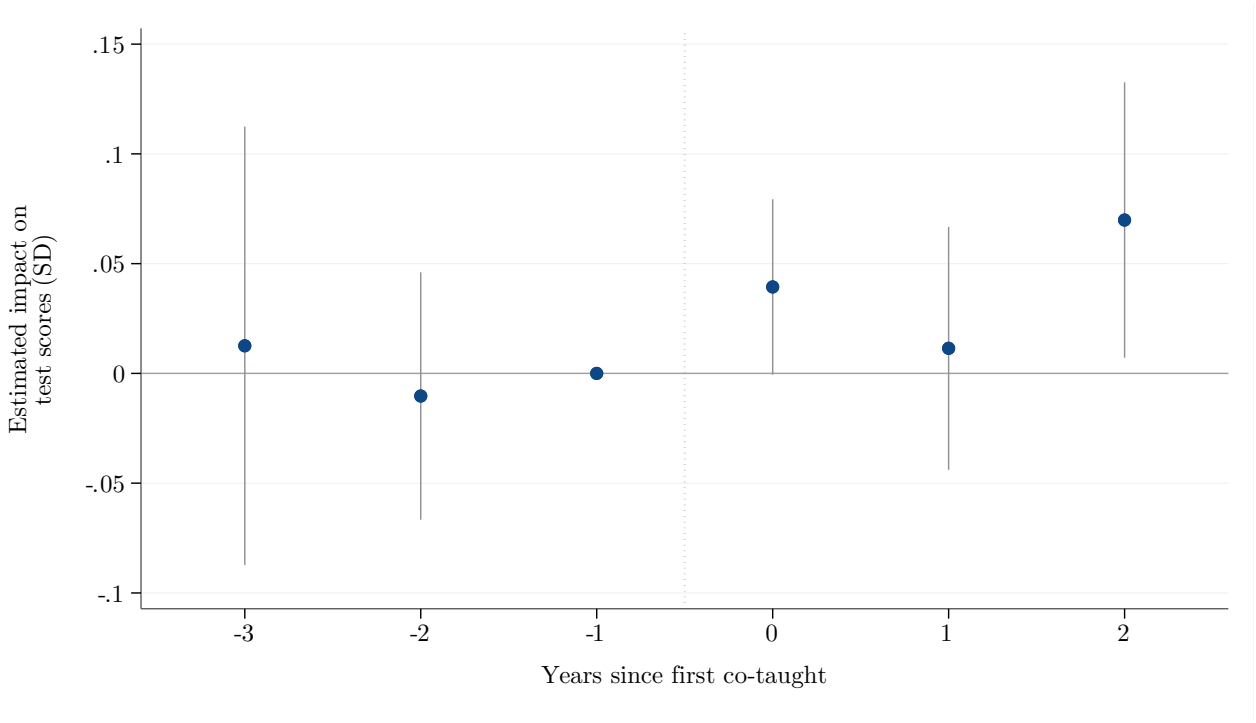
Rescaling to the teacher value-added distribution using middle school estimates from Chetty et al. (2014b), an effect of this size is equivalent to moving a teacher from the 50th to the 63rd percentile. That implies approximately $20,000 in undiscounted lifetime earnings per student, in 2025 dollars.

5.3 Gains Come From Experienced Partners

Partner experience explains most of the average effect. Teachers paired with highly experienced partners (16+ years) show gains of approximately 0.10σ (Figure 6). That is

[4] Post-period outcomes are observed only for teachers who stay, so these estimates condition on survival. A doubly robust risk-set design, which holds treated and never-co-teaching teachers to the same survival requirement, finds no retention difference between them (−2.2 and −3.7 percentage points at one and two years, neither distinguishable from zero). Because those estimates are imprecise enough to admit differences of either sign, I bound the achievement effect under selection in both directions, trimming the treated arm in one case and the controls in the other; the two-year effect stays within $[0.06,\ 0.08]\sigma$. Appendix B reports both exercises.

Figure 5: Difference in Student Test Scores By Time Since First Co-taught



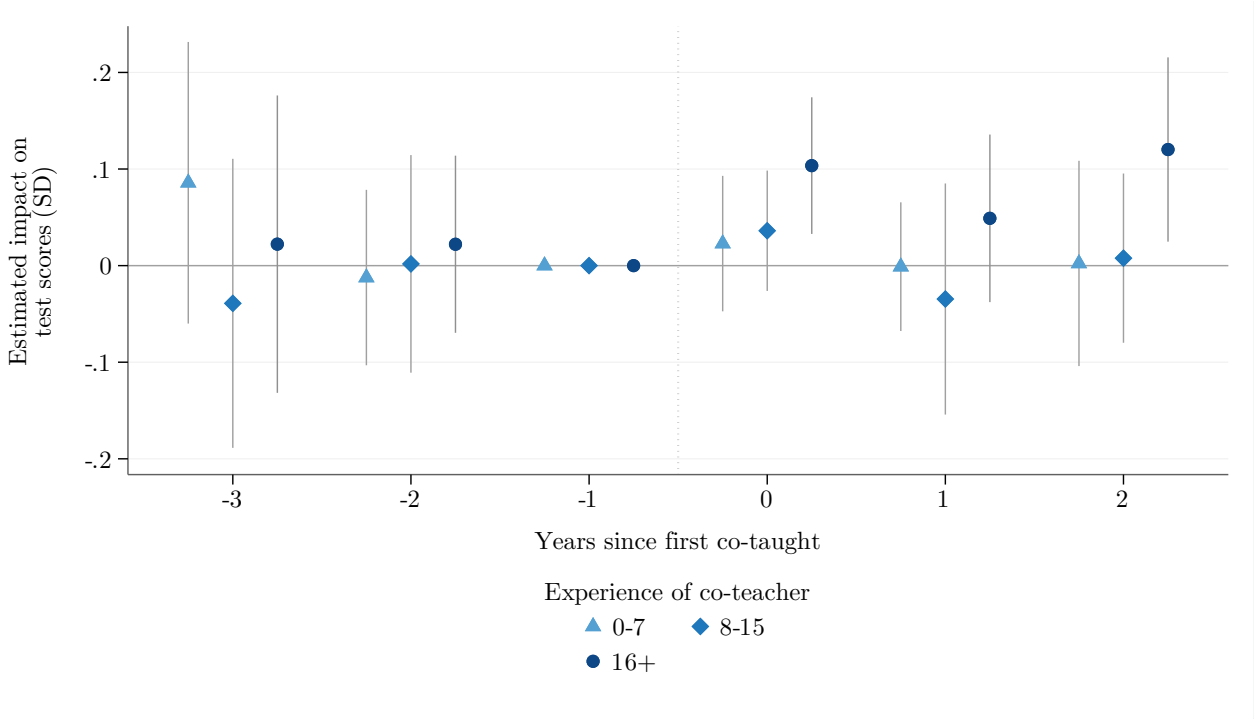
This figure presents stacked event study estimates of the effect of co-teaching on student test scores (standardized by grade and year). Each point plots $\hat{\omega}_k$ from Equation (3), representing the difference in test scores between teachers who co-taught and matched never-co-taught controls at event time k, where $k = 0$ is the first year of co-teaching. The omitted category is $i = -1$ (one year before first co-teaching). The sample includes general education teachers in Indiana from 2012–2019 with 1–3 years of solo experience before first co-teaching, matched within event groups defined by subject, cohort year, co-teaching start year, and exit year. Controls and fixed effects are the full set given in Equation (3): subject-interacted cubic polynomials in the previous two years' math and ELA scores, classroom composition shares, event-by-teacher and event-by-year effects, class size, grade by subject, student demographics and status-transition paths, and school-by-free-lunch-status effects. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

equivalent to moving a teacher from the 50th to the 81st percentile of the value-added distribution, implying approximately $50,000 in undiscounted lifetime earnings per student (Chetty et al., 2014b). Teachers paired with inexperienced partners (0–7 years) show no benefit.

The two arms come from separate regressions that share the never-co-taught controls, so their estimates are correlated and cannot be compared by eye. A pooled specification that interacts co-teaching with partner-experience bins supplies the missing covariance. In the primary sample the gap between the 16+ and 0–7 arms is too imprecisely estimated to distinguish from zero, which is what 141 treated teachers spread across three bins will support. The imputed-entry sample of Section 6 carries 393 treated teachers, and there a gap of the same size, 0.066 standard deviations, is estimated precisely enough to reject equality

($p = 0.036$); partner experience entered continuously gives the same answer ($p = 0.076$). The contrast does not change between samples, only the standard error, which is the signature of a power constraint rather than a fragile result. Pre-period coefficients are jointly indistinguishable from zero in both arms of that sample ($p = 0.61$ for 16+ partners, $p = 0.89$ for 0–7), so the contrast is not inherited from divergent trajectories before entry.

Figure 6: Difference in Student Test Scores By Time Since First Co-taught and Co-Teacher Experience



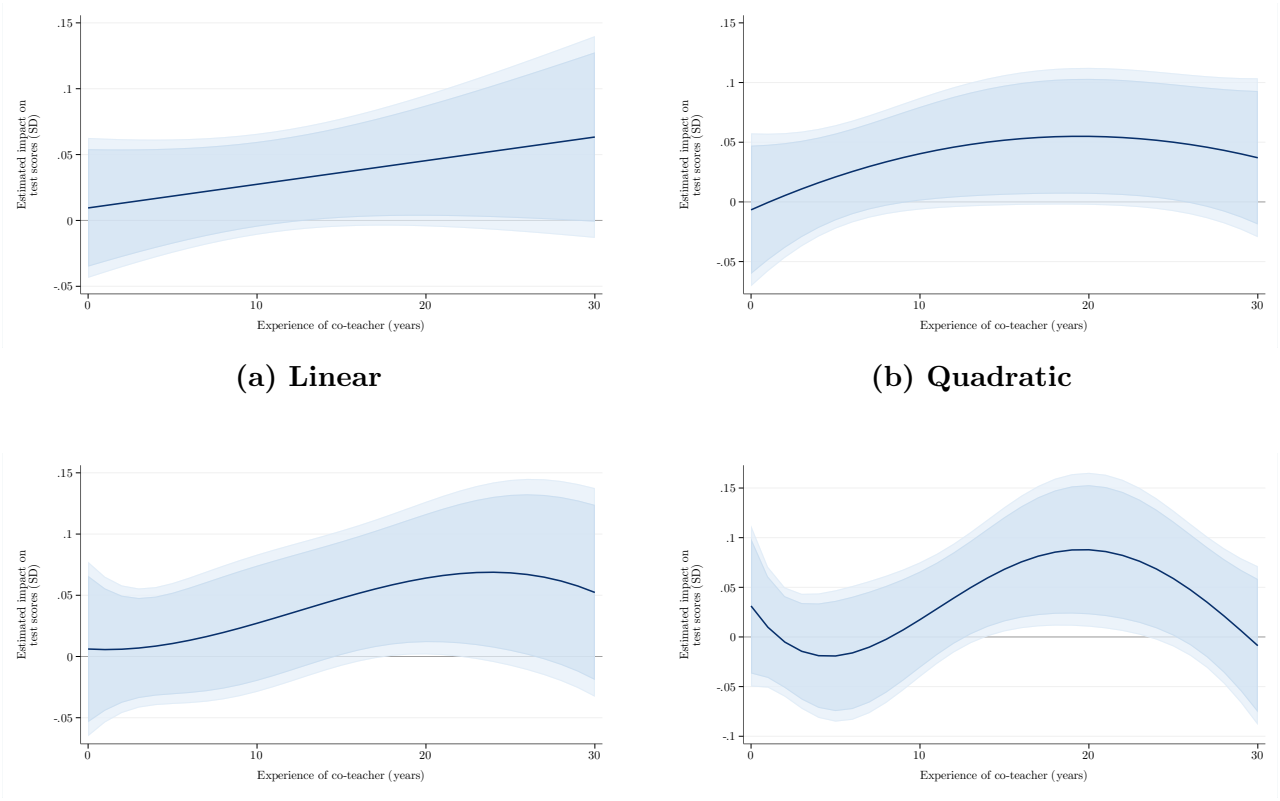
This figure disaggregates the main event study (Figure 5) by the experience level of the co-teaching partner: 0–7 years, 8–15 years, and 16+ years. Each panel plots stacked event study estimates from Equation (3) estimated separately for each partner experience subgroup. The sample, controls, and fixed effects are identical to those in Figure 5. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

I estimate separate difference-in-differences for finer experience bins to identify where benefits are largest. Benefits increase with partner experience and then decline at the highest levels (Figure A.10).[5] The pattern holds under alternative functional form assumptions (Figure 7).

The data do not reveal why benefits peak in the mid-career range and decline among the

[5]Because the separate by-bin regressions share the same never-co-taught controls, comparing bins to one another requires a pooled specification that interacts co-teaching with five-year partner-experience bins. There the 25+ effect sits below the 15–19 peak ($p = 0.03$), though a joint test does not reject equality across all bins ($p = 0.16$). I therefore read the location of the peak as suggestive.

Figure 7: Co-Teacher Experience Effects Under Alternative Functional Forms



This figure presents the relationship between co-teacher experience and post-co-teaching student test score improvements under alternative polynomial specifications (linear through quartic). Each panel interacts a post-co-teaching indicator with a polynomial of co-teacher experience. Shaded regions represent 90% (darker) and 95% (lighter) confidence intervals. All specifications use the stacked difference-in-differences design with the event study sample (general education teachers with 1–3 years of prior solo experience), and the full control and fixed-effect set of Equation (3), with standard errors clustered at the teacher level.

most experienced partners. Pedagogical practices evolve, so teachers who developed their approaches decades ago may transfer less relevant techniques. Alternatively, very experienced teachers may have internalized their practice to the point where they struggle to articulate what they do.

5.4 Interpreting the Partner Experience Gradient

The relationship between partner experience and post-co-teaching gains is the paper's central result. Knowledge transfer from experienced to inexperienced teachers is the simplest explanation for it, though I cannot observe what transfers, and the rest of this subsection works through the alternatives.

First, changes in student assignment are unlikely to explain the results. Co-teachers receive more students with disabilities and from disadvantaged backgrounds, which biases estimates downward. If unobserved student sorting were driving the results, weakening the student controls should make the effect appear. It disappears instead (Figure A.12).

Second, reduced workload during co-teaching cannot explain gains that persist after the partnership ends, when teachers are back in solo classrooms with full instructional responsibility. Some post-period years still include active co-teaching, but controlling for current co-teaching status leaves the results unchanged (Figure A.9). Dropping year zero, which is co-teaching by construction, leaves the two-year estimate significant (Appendix Table A.7).

Third, co-teaching might improve outcomes through motivation, confidence, or reduced burnout from a supportive partnership rather than through learning. Any partner could supply those, so this channel predicts similar benefits regardless of partner experience. The estimates instead rise with partner experience, reaching 0.10σ for 16+ partners (Figure A.10).

Fourth, what a partner has to offer could come from formal training rather than accumulated practice. Teachers with master's degrees or doctorates are no more beneficial as partners than those with bachelor's degrees (Figure A.11). Only classroom experience predicts gains.

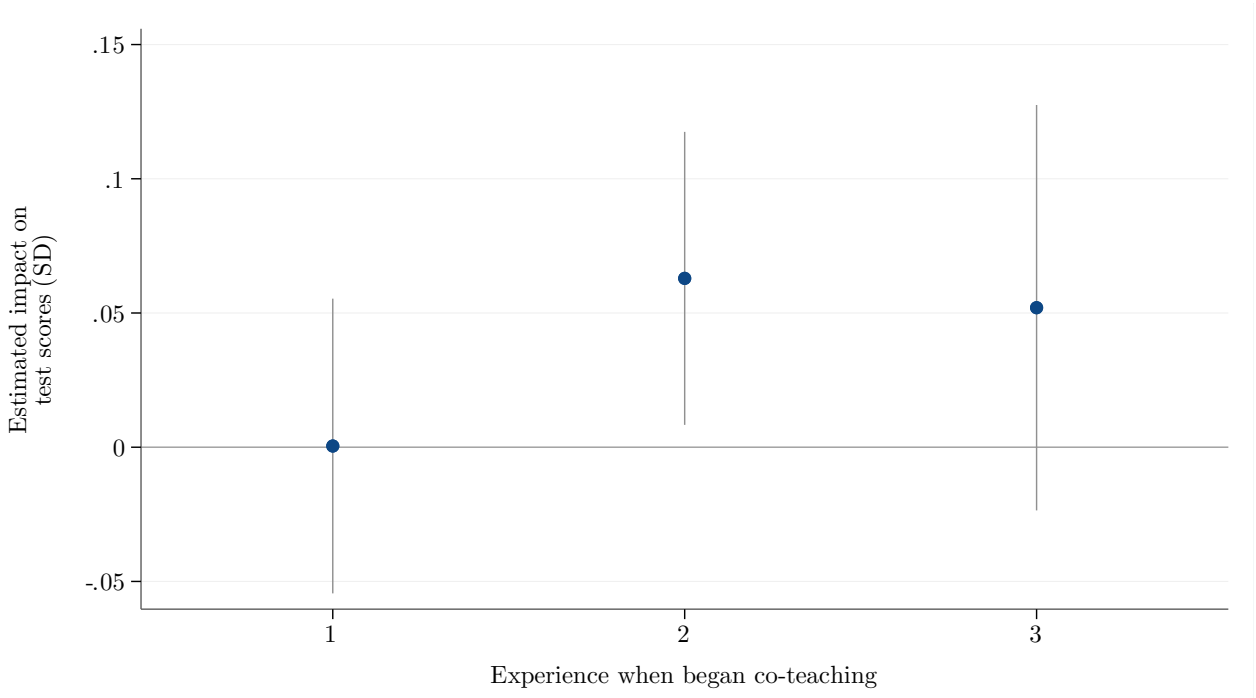
Two benchmarks discipline these magnitudes. Intensive teacher-coaching programs assign teachers a trained coach for periodic one-on-one sessions and raise student achievement by roughly 0.18σ (Kraft et al., 2018). The 16+ partner effect of 0.10σ is around half that, from an arrangement that delivers a full year of daily collaboration. Against the returns-to-experience profile, teachers gain roughly 0.064σ over their entire first five years (Papay and Kraft, 2015). The 16+ estimate is larger, though the two are not measuring the same thing: one is an average developmental path, the other is conditional on drawing an unusually experienced partner. These are large effects, but they fall within the range that intensive professional development achieves, and daily co-teaching is more intensive than anything in that literature. Where an individual estimate runs higher, as with the

imputation estimator's year-1 point estimate of 0.463σ, imprecision rather than the design is responsible: its confidence interval comfortably includes the stacked estimates. The average effect of 0.04σ is modest by any benchmark.

5.5 Teachers With Some Solo Experience Benefit Most

The benefit also depends on where the teacher is in their own career. Teachers who began co-teaching with two years of prior solo experience gain the most (0.063σ, s.e. 0.028). The three-year estimate is similar but imprecise (0.052σ, s.e. 0.038). Teachers in their first year gain nothing (0.000σ, s.e. 0.028; Figure 8).

Figure 8: Difference in Student Test Scores By Experience When First Co-taught



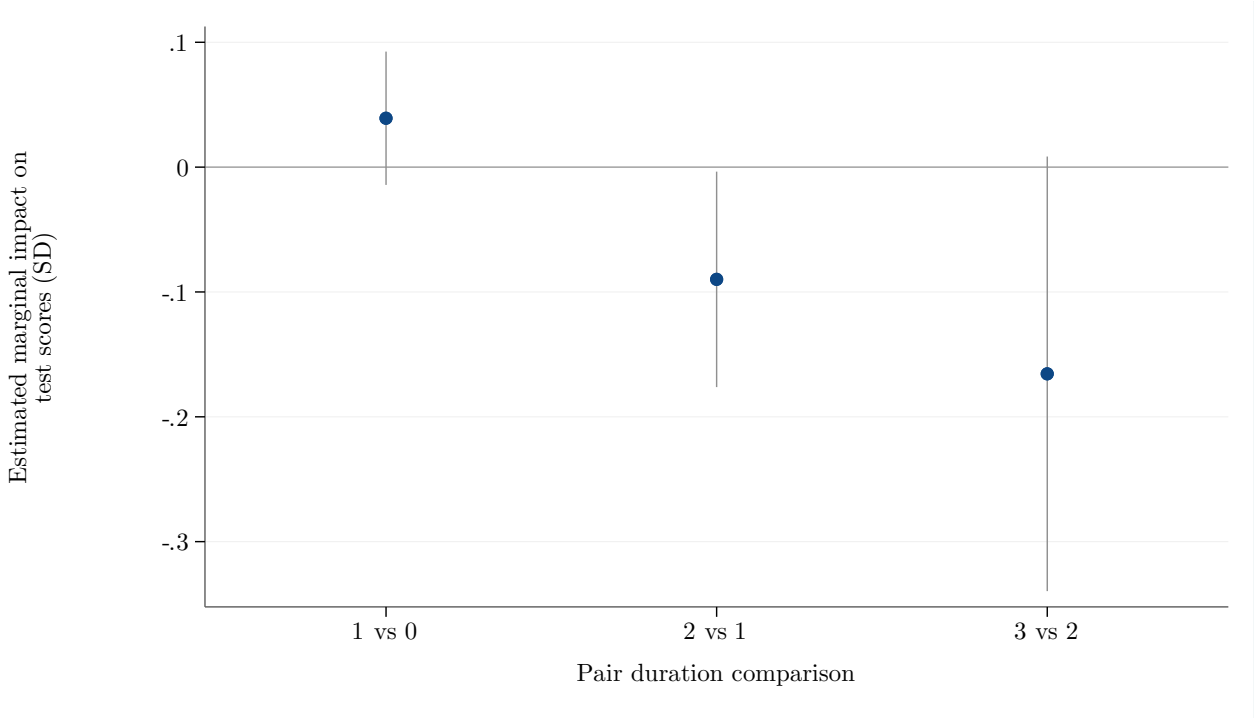
This figure estimates the effect of co-teaching on post-co-teaching student test scores by the teacher's years of solo experience at the time they first co-taught (1, 2, or 3 years). Each point represents a separate stacked difference-in-differences estimate. The sample, controls, and fixed effects are identical to Figure 5. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

Some baseline competence may be needed to recognize and absorb what an experienced partner does. First-year teachers may also be too occupied with basic classroom management to attend to subtle pedagogical techniques. I cannot distinguish these explanations.

5.6 One Year of Co-Teaching Is Enough

Additional years of co-teaching do not add to the benefit, and the estimates suggest they subtract from it. I test whether skill transfer accumulates by estimating the marginal effect of each additional year of co-teaching (Figure 9). Each point is a separate stacked difference-in-differences estimate comparing teachers who co-taught exactly n years to those who co-taught exactly $n-1$ years. The controls and event-specific fixed effects follow Equation (3), so each coefficient isolates the incremental value of the nth year.

Figure 9: Marginal Effect of Additional Years of Co-teaching



This figure estimates the marginal effect of each additional year of co-teaching on post-co-teaching student test scores. Each point compares teachers who co-taught for exactly n years to teachers who co-taught for exactly $n-1$ years, using stacked difference-in-differences. The sample, controls, and fixed effects follow the same specification as Figure 5, with the comparison group varying by duration. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

The first co-teaching year carries a positive marginal effect (0.04σ, s.e. 0.03). The second year is negative (-0.09σ, s.e. 0.04), and the third more so (-0.17σ, s.e. 0.08). The third-year comparison rests on only 66 pairs. Comparisons beyond the third year are too thin to estimate.

Extended co-teaching may come at the expense of solo teaching experience. A teacher in

a third year of co-teaching spends another year observing a partner rather than being forced to solve problems independently. If the benefit of collaboration comes from observing a partner's techniques rather than from cumulative practice together, the teacher with more solo experience may develop faster. Section 4 points the same way: pairs do not improve together over time once general experience is controlled.

6 Robustness

6.1 Prior Student Achievement

Conditioning on students' prior achievement is the standard way to measure teacher productivity, because it strips out the part of measured output that comes from which students a teacher was assigned. Chetty et al. (2014a) show that value-added measures built this way are not meaningfully biased by student sorting. Dropping the controls shrinks the estimate (Figure A.12), which is what assignment to students with lower expected growth would produce. Demographic composition need not shift for this to happen, since prior scores capture student difficulty more finely than the shares in Table 2. The specification without prior-score controls is therefore a biased benchmark. Removing the controls also changes the estimand from growth to levels, so the comparison alone does not settle the point.

The result does not hinge on how I implement this control. Using one year of lagged scores rather than two changes little (Figure A.13), and varying the polynomial in prior scores from linear through quartic leaves the estimate essentially unchanged (Figure A.14). Nor does the timing of co-teaching entry coincide with a discrete shift in who is assigned to the classroom. Estimating the event study with student characteristics as outcomes shows no sharp change in prior achievement, English learner status, or the share of students of color at entry (Figure A.15). The sorting documented in Table 2 is a level difference the prior-score controls absorb, not a dynamic confound.

I also consider student fixed effects, which absorb every fixed student characteristic. A

student fixed effect has no identifying variation in the stacked design as constructed: within an event, treatment timing varies only between treated and control teachers, and students do not switch between them. The fix is to augment each event with its own students' observations from other teachers' classrooms, including veteran teachers outside the event-study population. These outside observations carry their own event-time category and event-specific teacher effects, so they contribute nothing to the treated-control contrast. Their only role is to pin down each student's fixed effect from that student's full test history. The coefficients are then estimated within teacher, within year, and within student, with the student effect replacing the prior-score controls, since including both would difference a lagged outcome against a unit effect. Each student's anchor years draw on the full tested-grade history, grades 3 through 8, with post-co-teaching years of any co-teaching anchor teacher absorbed by event-specific indicators.

The pattern survives this demanding specification, attenuated (Appendix Figure A.16). On 1,079,743 observations the pre-period coefficients are statistically zero and the post-period estimates are positive, 0.024σ (s.e. 0.019) in year 0 and 0.053σ (0.027) in year 2, with the 16+ partner subsample sitting above the average (0.049σ and 0.075σ) though imprecisely estimated. Attenuation relative to the main specification is partly mechanical, since differenced and fixed-effects specifications amplify measurement error in test scores. But measurement error need not be the whole story. If schools sort students to teachers on characteristics the lagged scores miss, part of the gap reflects sorting rather than noise, and the student fixed effects estimates are then a conservative lower bound. Little of the gap can plausibly be sorting: the dynamic composition checks above show no shift in observables at entry, and the far better-powered unstacked comparison below attenuates only modestly rather than by half. I treat the stacked student fixed effects estimates as a complementary check whose lower bound still lies above zero, not as the preferred specification.

A pooled comparison on the unstacked panel of all teachers' classrooms makes the same point with far more power (Appendix Table A.6). On 2.9 million student-year observations with identical samples and fixed effects, swapping the lagged-score controls for student fixed

effects moves the post-co-teaching coefficient from 0.050σ (s.e. 0.019) to 0.047σ (0.015). The 16+ partner estimate moves from 0.098σ (0.033) to 0.075σ (0.024). The experience gradient survives the alternative control strategy with precision to spare. The same regressions carry a placebo: post-co-teaching years of co-teaching teachers *outside* the analysis sample, whose entry timing is not measured cleanly, show no effect.

6.2 Sample Restrictions

I restrict the main analysis to teachers with 1–3 years of experience before co-teaching. The restriction follows from how teachers come to co-teach at different career stages. Early entry is the institutional default, while entry after four or more years is selected on classroom trajectories in ways an event study cannot absorb. This section works through that evidence. The headline experienced-partner gradient holds with or without the restriction; what the restriction sharpens is the pooled average.

The restriction also keeps the estimand coherent. For an early-career teacher, co-teaching is a candidate input into *development*, in the steep part of the returns-to-experience profile where skill formation is concentrated (Papay and Kraft, 2015). That is also the population a pairing policy would target. For a veteran whose entry follows a relative decline, the same arrangement is plausibly a response to circumstances, and its effect answers a different question. Pooling the two does not generalize the estimate; it averages a development effect with whatever mixture of reversion and response the late entrants' trajectories produce.

Co-teaching entry is concentrated in the first years of a career and falls steeply with experience (Figure 1). Junior teachers lack the seniority to refuse assignments that enrollment and scheduling generate, so early entry happens *to* teachers more than it is chosen by them.

I next test what predicts entry. Classroom performance does not. Neither the level of a teacher's classroom scores nor their year-over-year change predicts beginning to co-teach the following year, for either experience group, with point estimates under one percentage point

per standard deviation of classroom performance. I can rule out effects larger than roughly two percentage points against entry rates of a few percent. Nor are late entrants matched to systematically more experienced partners, as a mentoring-intervention story would predict (difference of 0.7 years, s.e. 1.6).

Late entry is, however, far more *predictable* from classroom circumstances. A propensity model on classroom characteristics discriminates veteran entry from non-entry with an AUC of 0.72, against 0.59 for junior entry, barely better than chance. What distinguishes late entry most sharply is its *trajectory*. Within treated teachers, late entrants' scores in the year before co-teaching sit 0.16σ below their own earlier years relative to early entrants, an Ashenfelter-dip pattern. The dip concentrates where the headline gradient lives: among late entrants paired with 16+ experience partners it is 0.27σ (s.e. 0.16).

The same split shows up in the event study's pre-trends. The pooled main specification's pre-period coefficients are jointly indistinguishable from zero ($F = 0.25$, $p = 0.78$). Rebuilding the stacked event study separately by entry band, the same test is clean for early entrants ($F = 0.22$, $p = 0.80$) but larger and sign-switching for late entrants ($F = 1.41$, $p = 0.25$). No single contrast is decisive, since the late-entrant cell is small, but they point the same way. Late entry is not random with respect to classroom trajectories, even though it is unrelated to performance levels. Selection on dynamics rather than levels is what an event study cannot tolerate and what level-based tests cannot catch.

A further check asks how good a counterfactual the never-co-teaching controls are, band by band. I rebuild the stacked design separately for each entry group and estimate treated-control gaps at each pre-period event time, in the outcome and in 11 covariates. For the 1–3 sample the design passes cleanly. The outcome's pre-period gaps are 0.01σ in magnitude and jointly indistinguishable from zero ($F = 0.22$, $p = 0.80$), and no covariate rejects balance at the five percent level. For the 4+ sample the pre-period outcome movements are several times larger and non-monotone, covariate balance begins to fail, and entry itself is predictable. The controls are a valid counterfactual where the design uses them.

Three follow-on checks sharpen the picture. Entry into co-teaching often coincides with a

school move (Appendix D), so I re-estimate the main event study excluding treated teachers who changed schools in their first co-teaching year. The estimates are essentially unchanged (year 0 of 0.041σ versus 0.039σ; year 2 of 0.064σ versus 0.070σ), so relocation shocks bundled with entry are not generating the gains.[6]

The headline partner-experience gradient does not depend on the sample restriction. Estimated on all entrants with one or more years of prior experience, the 16+ partner event study yields 0.104σ (s.e. 0.032) in year 0 and 0.152σ (0.046) in year 2, nearly identical to the main sample.

The late-entrant version of that gradient is where selection bites. Estimated on late entrants alone, the 16+ event study is noisy with unstable pre-period coefficients, and the pre-entry dip concentrates in those same 16+ pairings. Veterans who begin co-teaching alongside an experienced partner tend to have been declining beforehand, so their post-period recovery mixes skill transfer with reversion. This is a small group, and it is excluded from the main analysis. The gradient holds whether or not they are included, but excluding them makes the estimate easier to interpret.

A final check turns to the known-first-year restriction itself. I impute each teacher's entry year from the administrative experience field and rebuild the event study on the unrestricted panel. Imputed-entry cohorts run back to 2006, with no restriction on how long after entry co-teaching begins, which recovers the veteran teachers the 2012 panel start drops: 393 treated teachers, 246 of them with no known first year, roughly 2.8 times the main sample. On this broad, veteran-inclusive sample the experienced-partner gradient persists and stays significant: 0.056σ (s.e. 0.026) in year 0 and 0.103σ (0.030) in year 2 (Figure A.18). The gain from an experienced partner is not an artifact of the recent-entry sample.

What the expansion dilutes is the pooled *average*, which attenuates toward zero (0.019σ here, 0.013σ under the stricter administrative-clock rebuild). This is the contamination this

[6] Co-teaching sections are created in response to special-education enrollment. Within a school-grade cell, an additional special-education student raises the probability the grade is co-taught, concentrated at the extensive margin (Appendix Figure A.1). Which teacher is pulled in is idiosyncratic conditional on the cell: classroom performance, partner experience, and pre-entry trends do not predict entry. The enrollment that creates a co-taught section is unrelated to the focal teacher's quality.

subsection documents rather than a fragility of the effect. The recovered returners enter with declining pre-period trajectories: restricting treated teachers to them alone drives the year -3 pre-period coefficient to -0.50σ (s.e. 0.15). Pooling them in averages a development effect with reversion. The restriction sharpens the average; the gradient, identified off partner experience rather than entry timing, holds without it.

A complementary check holds the known-first-year eligibility criterion fixed and varies which years the panel must be balanced on. The baseline design requires a treated teacher to have a solo-taught classroom in every event year, so a teacher missing one in any single year drops out entirely. Two features of the panel make that demanding: horizons beyond year two are unobservable, and in year zero the teacher is still partnered. I therefore sweep balanced-subset specifications, toggling which pre-periods, year zero, and post-horizons must be balanced, and keeping the panel balanced within whichever years each specification requires (24 distinct specifications; Appendix Table A.7). Each requirement admits a different subset of teachers, so the sweep varies the estimation sample as well as the construction. The experienced-partner effect is stable throughout: the two-year post-solo estimate has a median of 0.18σ and is significant in all 16 specifications that reach that horizon. Individual pre-period placebo coefficients occasionally cross significance, with 4 of 18 doing so at the five percent level. None survives a Benjamini–Hochberg correction for the multiplicity of placebo tests, so the apparent pre-period movements are consistent with sampling noise rather than selection.

Role classification provides a final sample-definition check. The paper follows general education teachers, but the administrative special-education setting codes are not fixed teacher attributes. Some teachers switch, most likely because a general education teacher moves into special education or the reverse, and 104 of the 141 treated teachers carry a general education code in every observed year. I rebuild the event study with both arms restricted to these consistently classified teachers. The year-0 estimates hold up, at 0.042σ (s.e. 0.023) on average and 0.081σ (0.037) for 16+ partners. The year-2 estimates attenuate and lose precision (0.018σ, s.e. 0.036 on average; 0.034σ, s.e. 0.067 for 16+), which is the

expected cost of discarding a quarter of the treated arm and most of the longer post-period horizons with it. Because a role switch may itself follow from co-teaching, this restriction conditions on a post-treatment outcome, so I read it as a check on classification rather than as a preferred specification. The near-term gradient does not depend on the mixed-role teachers; the year-2 estimates lean on them for precision.

Consistent with contamination rather than fragility, including all teachers with at least one year of prior experience attenuates the estimated benefits but preserves the pattern (Figure A.17).

The post-period coefficients pool teachers in different states: after year zero, some treated teachers are no longer partnered while others still co-teach. Beyond conditioning on current co-teaching status (Figure A.9), splitting treated teachers by spell length isolates the clean case, reported in Section 5.

6.3 Estimator, Weighting, and Trend Sensitivity

Three further checks are reported in Appendix C. The stacked design's implicit event weights, alternative weighting schemes, and the Borusyak et al. (2024) imputation estimator all deliver the same average effect and the same partner-experience gradient. Relative-magnitudes bounds (Rambachan and Roth, 2023) show the gradient survives meaningful violations of parallel trends while the pooled average, a borderline estimate, does not.[7] And building the control set up sequentially shows the prior-score cubics doing the work, with later layers moving the estimates by thousandths of a standard deviation.

[7]The robust confidence set for the year-0 16+ effect excludes zero for violations up to $\bar{M} = 0.75$, and for the pooled post-period average up to $\bar{M} = 0.25$, where $\bar{M}$ scales the largest admissible post-period violation of parallel trends against the largest one observed in the pre-period. Both sit below the conventional benchmark of one, which is a more demanding standard here than the ratio suggests: the pre-period anchor is roughly 0.02σ, so even $\bar{M} = 0.75$ admits only about 0.015σ of drift per year.

7 Policy Simulation

Pairing early-career teachers with highly experienced colleagues produces lasting productivity gains, but implementing such a policy involves costs. The experienced teacher must leave their solo classroom, and a less effective instructor may take their place. I simulate this tradeoff to determine when strategic pairing yields net positive effects on student achievement.

The policy pairs each early-career teacher with a 16+ year mentor for one year, after which both return to solo instruction. The development effect ($\delta = 0.10\sigma$) comes from my main estimates and the opportunity cost ($\gamma = 0.14\sigma$) from the novice-veteran productivity gap in Papay and Kraft (2015). A third channel, the contemporaneous benefit to students sharing a room with two teachers ($\beta = 0.02\sigma$), is taken from Jones and Winters (2024) rather than from my own design, which measures solo classrooms only; scenarios 2 and 3 set it to zero. No scenario credits a retention benefit, so the mentor pool never expands beyond what 0.20 annual turnover supplies. Appendix E gives the steady-state mechanics and the full parameter set.

Two features drive the arithmetic. Costs land immediately and once: a mentor leaves a solo classroom for a single year. Benefits accrue slowly but repeatedly, since a teacher who gains skill carries it to every later cohort. That asymmetry is what makes the policy viable at all, and it is why the useful summary is not the point estimate but the break-even opportunity cost γ^*, the value of γ at which the policy stops paying for itself.

Table 4 reports three scenarios that progressively remove optimistic assumptions: the baseline, then dropping the contemporaneous effect, then adding 10% annual decay of the acquired skill.

Mentor supply binds. With retention flat, 2.9% of teachers carry 16+ years of experience in steady state and 14.4% are ever treated, so the policy reaches a minority of the workforce however well it is targeted. The net annual effect runs from 0.004σ to 0.008σ per classroom, worth \$2,093 to \$4,021 in undiscounted lifetime earnings per student per year of exposure.

Table 4: Policy Simulation Results Across Scenarios

	(1) Baseline	(2) No Contemp. Effect	(3) Most Conservative
Panel A: Assumptions			
Contemporaneous effect	Yes	No	No
No skill decay	Yes	Yes	No
Panel B: Annual Effects (average school wide)			
Treatment effect on VA	$+0.011\sigma$	$+0.011\sigma$	$+0.008\sigma$
Contemporaneous benefit	$+0.001\sigma$	—	—
Opportunity cost	-0.004σ	-0.004σ	-0.004σ
Net annual effect	$+0.008\sigma$	$+0.007\sigma$	$+0.004\sigma$
Lifetime earnings per student	+\$4,021	+\$3,734	+\$2,093
Panel C: Cost-Benefit Analysis (per classroom)			
Cost	\$1,723	\$1,723	\$1,723
Societal benefit	\$80,422	\$74,678	\$41,852
Government benefit (15%)	\$12,063	\$11,202	\$6,278

Notes: This table presents steady-state effects of strategic teacher pairing under three scenarios. No scenario credits a retention benefit, so the mentor pool is fixed at 2.9% of teachers and 14.4% are ever treated in all three; those quantities are properties of the calibration rather than of the scenarios and are not shown separately. Column (1) includes the development and contemporaneous effects; column (2) removes the contemporaneous effect; column (3) adds skill decay. Skill decay (column 3 only) means the 0.10σ development effect diminishes by 10% each year post-treatment. Annual effects are measured per average classroom. Lifetime earnings per student are undiscounted gains in 2025 dollars per year of exposure, calculated by converting the net annual effect to teacher value-added units and applying earnings estimates from Chetty et al. (2014b), inflation-adjusted from 2010 to 2025 dollars. Cost per classroom equals the mentor pool share times \$60,000 (annual teacher salary). Societal benefit equals lifetime earnings per student times 20 students. Government benefit assumes 15% of societal benefit is captured through taxation.

These are modest per-student amounts, bounded by mentor supply rather than by the size of the development effect. Cumulative effects turn positive within 6 to 8 years in every scenario.

The margin for error is the more robust result, because it does not require the point estimate to be right. Break-even opportunity cost ranges from 0.29σ to 0.42σ across scenarios, 2 to 4 times the Papay and Kraft (2015) estimate of 0.14σ. Pulling an experienced teacher out of a classroom would have to cost several times what the returns-to-experience literature implies before the policy stopped paying off.

Against costs, the accounting is favorable but small in absolute terms. The policy reassigns existing staff rather than expanding the workforce, but it is not free: a mentor who

co-teaches leaves a solo classroom that someone must cover, and at the margin that cover is a hire, so I price the mentor at a full $60,000 salary. That gives $1,723 per classroom against a societal benefit of $41,852 to $80,422, or $6,278 to $12,063 captured through taxation at 15%. The fiscal benefit-cost ratio runs from 3.6:1 to 7.0:1.

8 Conclusion

Workplace collaboration can produce lasting skill development. Teachers who co-taught with experienced colleagues show higher student achievement after returning to solo instruction. The effect averages 0.04σ overall and reaches 0.10σ when paired with highly experienced partners. Rescaling to the teacher value-added distribution (Chetty et al., 2014b), the average effect corresponds to moving a teacher from the 50th to the 63rd percentile, and the highly experienced partner effect to the 81st. Together these imply $20,000 to $50,000 in undiscounted lifetime earnings per student. This pattern is consistent with knowledge transfer from experienced to inexperienced teachers. Formal credentials do not predict learning gains. Only experience does.

Three findings inform collaboration design. First, partner selection matters. Pairing with experienced colleagues generates lasting benefits, while estimates for novice-novice pairings sit near zero. Second, collaboration can be brief. One year appears sufficient, and extended partnerships may reduce development by coming at the expense of solo experience. Third, timing matters. Teachers with 2–3 years of independent experience benefit most. Some baseline competence may be necessary to absorb what experienced partners offer.

These findings suggest that organizations fail to extract value from collaboration they already pay for. Schools do not underinvest in co-teaching, since federal mandates guarantee substantial spending on it. But current practice allows scheduling constraints rather than developmental objectives to shape partnership formation, leaving the return on that mandated spending unclaimed. The policy simulation indicates that pairing experienced workers with junior colleagues can generate annual gains of 0.008σ per classroom under the

baseline assumptions, a reallocation of existing personnel rather than a new expenditure. Benefits remain positive under conservative assumptions about skill decay and auxiliary effects, though the scale is limited: because only 14.4% of teachers can be paired with a 16+ mentor in steady state, the policy reaches a minority of the workforce however well it is targeted.

Several limitations bound the interpretation. The experienced-partner gradient is the paper's robust finding, surviving alternative estimators, control sets, worst-case attrition trimming, and the expansion to veteran co-teachers. The pooled average is estimated less precisely, and some heterogeneity cells carry wide confidence intervals. Co-teaching is a specific form of collaboration that may not generalize to other settings. The estimates also depend on controlling for prior student achievement, standard in the teacher-productivity literature but a specification choice nonetheless. Knowledge transfer is the simplest explanation for the partner-experience gradient, but I cannot observe what transfers.

Collaboration's value may lie not just in what partners produce together, but in how working together changes what each can produce alone. Organizations that pair the right workers, early but not too early, for a short spell may accelerate workforce development. Current practice appears to leave many of these potential gains unrealized.

References

Andersen, S. C., Beuchert, L., Nielsen, H. S., and Thomsen, M. K. (2020). The effect of teacher's aides in the classroom: Evidence from a randomized trial. *Journal of the European Economic Association*, 18(1):469–505.

Arrow, K. J. (1962). The economic implications of learning by doing. *The Review of Economic Studies*, 29(3):155–173.

Austin, V. L. (2001). Teachers' beliefs about co-teaching. *Remedial and Special Education*, 22(4):245–255.

Azoulay, P., Graff Zivin, J. S., and Wang, J. (2010). Superstar extinction. *The Quarterly Journal of Economics*, 125(2):549–589.

Bessette, H. J. (2008). Using students' drawings to elicit general and special educators' perceptions of co-teaching. *Teaching and Teacher Education*, 24(5):1376–1396.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event-study designs: robust and efficient estimation. *The Review of Economic Studies*, 91(6):3253–3285.

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134(3):1405–1454.

Chen, Y. (2021). Team-specific human capital and team performance: evidence from doctors. *American Economic Review*, 111(12):3923–3962.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–2679.

Conderman, G. and Hedin, L. R. (2014). Co-teaching with strategy instruction. *Intervention in School and Clinic*, 49(3):156–163.

Cook, S. C., McDuffie-Landrum, K. A., Oshita, L., and Cook, B. G. (2017). Co-teaching for students with disabilities: A critical and updated analysis of the empirical literature. In *Handbook of Special Education*, pages 233–248. Routledge.

De Backer, L., Schelfhout, W., Simons, M., and Vandervieren, E. (2023). Student teachers' peer team teaching experiences from a quantitative perspective: Perceptions, profiles and transition probabilities. *Teaching and Teacher Education*, 135:104361.

de Grip, A. and Sauermann, J. (2012). The effects of training on own and co-worker productivity: Evidence from a field experiment. *The Economic Journal*, 122(560):376–399.

Gershenson, S., Lindsay, C. A., Papageorge, N. W., Campbell, R. A., and Rendon, J. H. (2023). Spillover effects at school: How Black teachers affect their White peers' racial competency. NBER Working Paper 31847, National Bureau of Economic Research.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Guryan, J., Kroft, K., and Notowidigdo, M. J. (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, 1(4):34–68.

Haggag, K., McManus, B., and Paci, G. (2017). Learning by driving: Productivity improvements by new york city taxi drivers. *American Economic Journal: Applied Economics*, 9(1):70–95.

Ichino, A. and Maggi, G. (2000). Work environment and individual background: Explaining regional shirking differentials in a large Italian firm. *The Quarterly Journal of Economics*, 115(3):1057–1090.

Jackson, C. K. and Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, 1(4):85–108.

Jones, N., Kaler, L., Markham, J., Senese, J., and Winters, M. A. (2025). Service delivery models: Impacts for students with and without disabilities. *Educational Researcher*, 54(3):141–152.

Jones, N. and 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.

Kraft, M. A., Blazar, D., and 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.

Mas, A. and Moretti, E. (2009). Peers at work. *American Economic Review*, 99(1):112–145.

Nix, E. (2020). Learning spillovers in the firm. IFAU Working Paper 2020:14, Institute for Evaluation of Labour Market and Education Policy.

Ost, B. (2014). How do teachers improve? the relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, 6(2):127–151.

Papay, J. P. and Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, 130:105–119.

Papay, J. P., Taylor, E. S., Tyler, J. H., and Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, 12(1):359–388.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *The Review of Economic Studies*, 90(5):2555–2591.

Shaw, K. and Lazear, E. P. (2008). Tenure and output. *Labour Economics*, 15(4):704–723.

Thompson, P. (2010). Learning by doing. In Hall, B. H. and Rosenberg, N., editors, *Handbook of the Economics of Innovation*, volume 1, pages 429–476. Elsevier.

Thompson, P. (2012). The relationship between unit cost and cumulative quantity and the evidence for organizational learning-by-doing. *Journal of Economic Perspectives*, 26(3):203–224.

Wing, C., Freedman, S. M., and Hollingsworth, A. (2024). Stacked difference-in-differences. NBER Working Paper 32054, National Bureau of Economic Research.

A Appendix

Table A.1: Teacher Characteristics by Co-Teaching Status and Population

	(1) Never Co-Taught	(2) Ever Co-Taught	(3) Difference (1)−(2)	(4) No Known First Year	(5) All Indiana
Mean experience	1.15	1.38	-0.23^{***}	13.20	10.84
	(1.25)	(1.24)	(0.04)		
General education	0.91	0.72	0.19^{***}	0.87	0.86
	(0.28)	(0.44)	(0.01)		
Ever switches gen-ed/sped role	0.00	0.06	-0.06^{***}	0.02	0.02
	(0.00)	(0.24)	(0.01)		
Master's degree or higher	0.12	0.12	0.00	0.52	0.44
	(0.33)	(0.33)	(0.01)		
Age (years)	31.19	31.21	−0.02	43.51	41.01
	(8.49)	(8.13)	(0.28)		
Female	0.75	0.77	−0.02	0.79	0.79
	(0.43)	(0.42)	(0.01)		
White	0.90	0.90	0.00	0.92	0.92
	(0.30)	(0.30)	(0.01)		
Black	0.06	0.06	−0.01	0.06	0.06
	(0.23)	(0.24)	(0.01)		
Hispanic	0.02	0.02	0.00	0.01	0.01
	(0.14)	(0.13)	(0.00)		
Teachers	3,425	1,175		23,685	29,628

Notes: Columns (1)–(3) compare teacher-level characteristics for teachers who ever versus never co-teach within the known-first-year development-analysis population (teachers first observed in their first year of teaching, 2012–2019; younger than the Indiana average by construction): standard deviations are in parentheses, and column (3) reports the difference in means with unequal-variance standard errors. Columns (4)–(5) report means for teachers with no observed first year and for all Indiana teachers in the panel, drawn from the full teacher panel before the known-first-year restriction. The columns therefore group as follows: columns (1) and (2) partition the known-first-year population, column (4) is its complement, and column (5) is the union of the two; columns are not additive across groups. Experience is averaged over observed years; race categories do not sum to one. In Panel C, standard deviations are omitted for binary student characteristics, whose dispersion is determined by the mean.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.2: Sample Selection for the Event Study

	Distinct teachers	
Restriction	All	General education
Treated arm (ever co-taught)		
In known-first-year panel	1,175	877
First-year flag consistent across records	880	667
Entered co-teaching 1–3 years after start	250	222
Survives cohort assembly and balanced window	141	128
Control arm (never co-taught)		
In known-first-year panel	3,425	—
Matches a treated cohort (entry × exit × subject)	663	—
Event study sample (teachers)	804	
Event study sample (students)	61,652	

Notes: Distinct teacher counts at each restriction that defines the event study sample, mirroring the sample construction in Section 5. The treated arm begins with every teacher observed co-teaching in the known-first-year panel and narrows to those whose "First Year" flag is internally consistent, who entered co-teaching one to three years after their first observed year, and who survive assembly into balanced entry-year × exit-year × subject cohorts. Because the development analysis follows general education teachers, the final column reports the same funnel restricted to them. Classification uses the administrative special-education setting codes, which can vary within teacher over time; Section 6.2 re-estimates the design restricted to teachers classified general education in every observed year. The control arm consists of never-co-taught teachers whose career profile exactly matches a treated cohort. The two arms combine into the 804 teachers and 61,652 students of the event study column in Table 1.

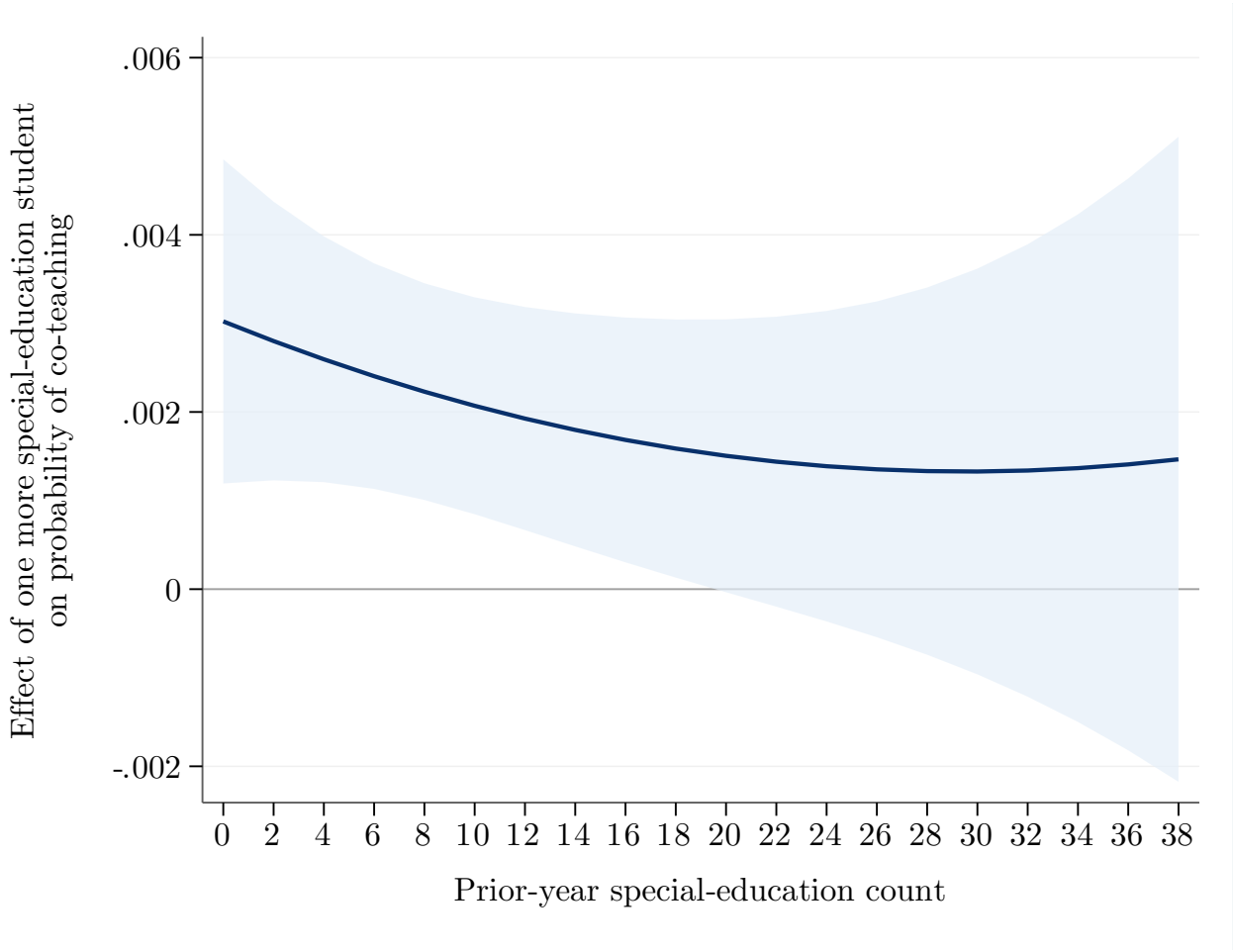
Table A.3: Assigned Partner Experience and Focal-Teacher Baseline Quality

	Partner experience (yrs) (1)	Partner is 16+ (2)
Pre-co-teaching student achievement	0.394	-0.034
	(0.975)	(0.050)
Teachers	136	136
R^2	0.001	0.003

Notes: Each column regresses a measure of the assigned co-teaching partner's experience on the focal general-education teacher's mean standardized student achievement in the pre-co-teaching period (event time −3 to −1), among treated teachers collapsed to the teacher level. The column 1 outcome is the partner's years of experience in the entry year; column 2 is an indicator for a partner with 16 or more years. A coefficient near zero indicates that experienced partners are not systematically assigned to teachers with stronger pre-co-teaching performance. The sample contains 136 of the 141 treated teachers in Table A.2: a teacher enters this regression only with observed student achievement in the pre-co-teaching window and observed partner experience in the entry year, and five treated teachers lack at least one of the two. Robust standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.1: Co-Teaching Responds to Special-Education Enrollment at the Extensive Margin

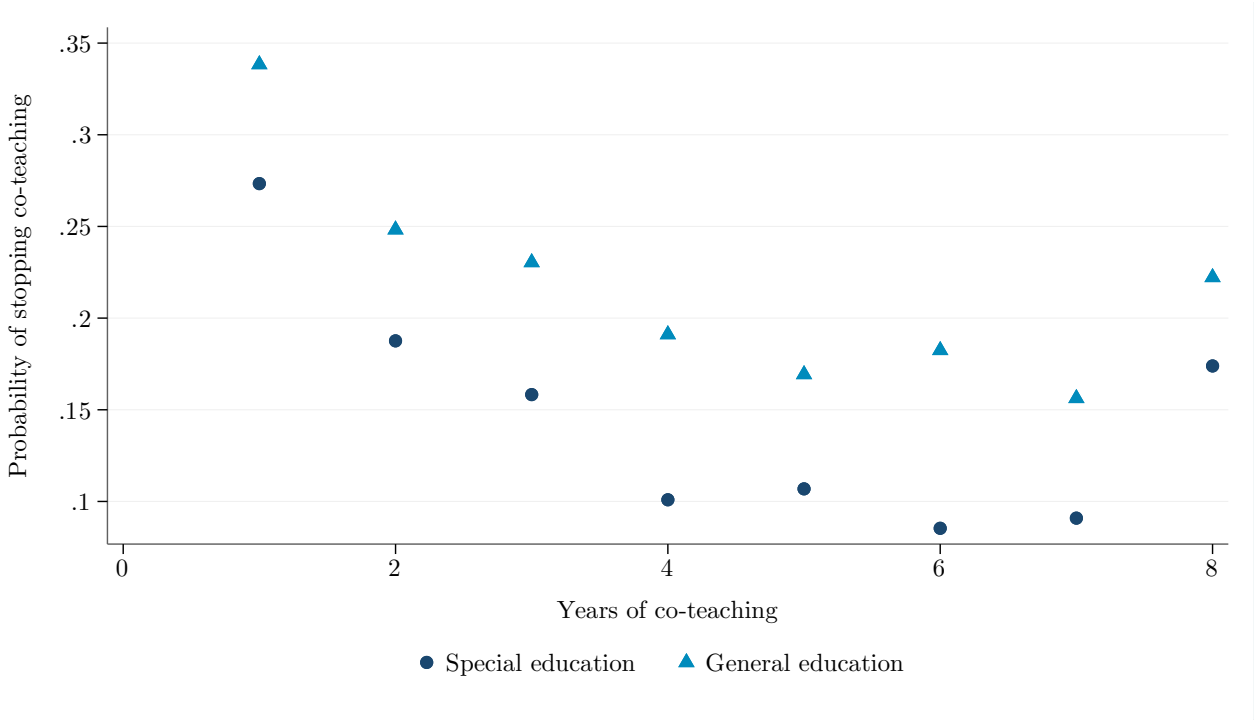


This figure plots the marginal effect of a one-student year-over-year change in a school-grade's special-education count on the probability the grade has a co-taught class, as a function of the grade's starting (prior-year) special-education count. The estimates come from a regression of co-teaching on the enrollment change interacted with a quadratic in the starting count, absorbing school-grade and year fixed effects, with standard errors clustered at the school level; the shaded band is the 95% confidence interval. The slope is positive and significant where a grade begins with few special-education students and declines toward zero as the starting count rises, indicating co-teaching is created on the extensive margin and saturates. The sample is all Indiana school-grade-years, 2012–2019.

B Attrition and the Post-Period Sample

Co-teaching could change who remains in the data, which would make the post-period estimates conditional on survival. I find no retention difference to speak of. A doubly robust risk-set design, comparing teachers who begin co-teaching in year g to observably similar never-co-teaching teachers drawn from the same risk set of teachers employed in g, puts the difference at -2.2 percentage points (s.e. 1.8) one year out, -3.7 (2.7) two years out, and 0.2

Figure A.2: Hazard Rate for Continuing Co-Teaching



This figure plots the annual probability that a teacher stops co-teaching in the following year, conditional on co-teaching in the current year and continuing to teach, separately by teacher type (general education and special education) and years of co-teaching experience. The sample includes all co-teaching teachers in Indiana from 2012–2019.

Table A.4: Characteristics of Short- and Long-Duration Co-Teaching Pairs

	(1) Pairs 1–2 years	(2) Pairs 3+ years
General education teacher experience	10.76	13.14
	(9.54)	(9.65)
Special education teacher experience	11.03	14.58
	(10.18)	(10.09)
Student test score	−0.52	−0.45
	(0.97)	(0.96)
Lagged student test score	−0.57	−0.53
	(0.94)	(0.91)
Pair-years	9,778	3,472

Notes: This table compares pair-year observations for co-teaching pairs whose completed duration is one to two years against pairs lasting three or more years. Experience is measured in years. Test scores are standardized by grade and year; means are taken over student-year observations linked to each pair-year where scores are observed. Standard deviations in parentheses. Pairs first observed in 2012 are excluded because their start year, and hence duration, is censored.

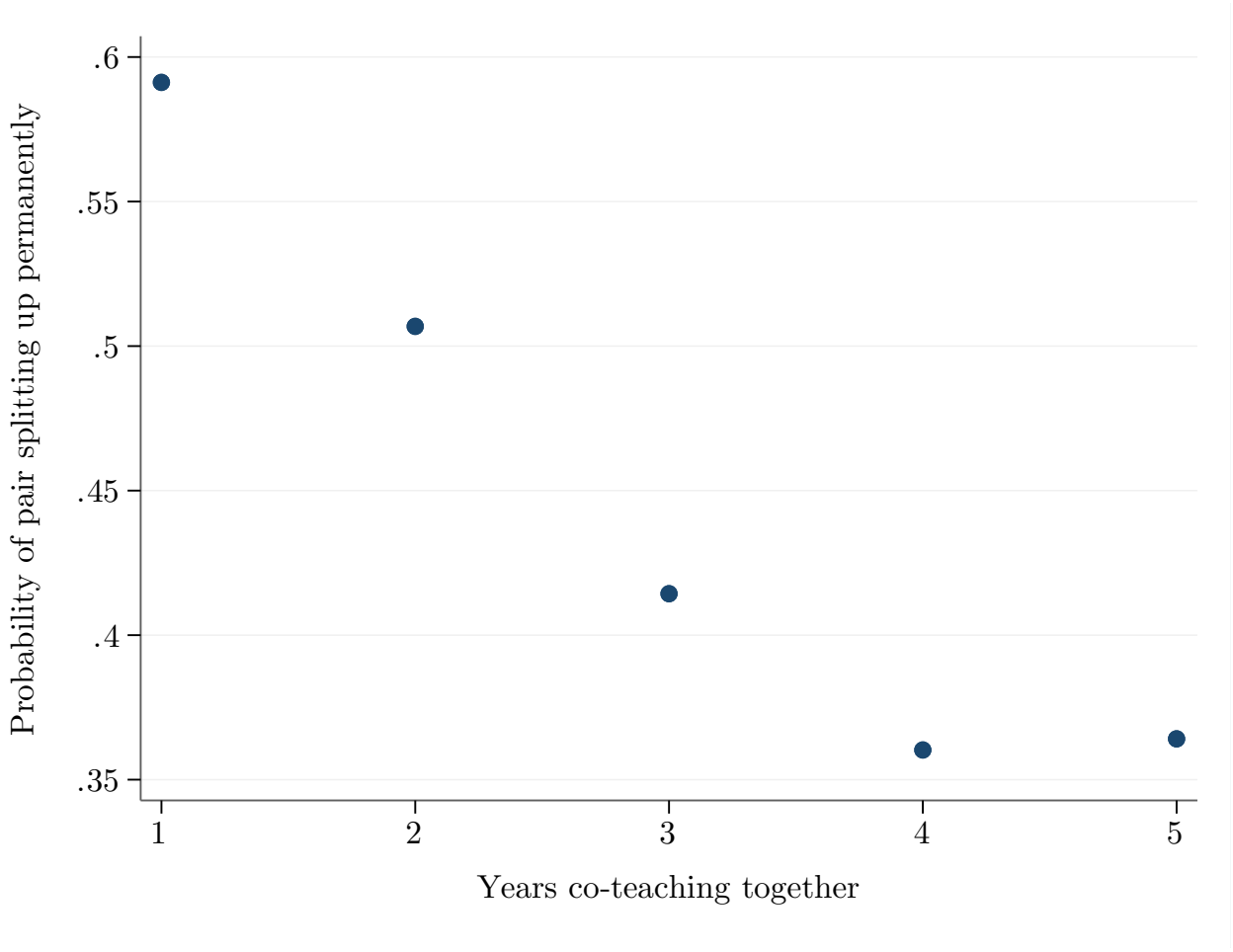
Table A.5: Pair-Tenure Profile under Common and Group-Specific Experience Profiles

	Adjusted Student Test Scores		
	Common profile (1)	Split at $e_0 \leq 2$ (2)	Split at median e_0 (3)
1 year together	−0.0248	−0.0252	−0.0162
	(0.0226)	(0.0229)	(0.0242)
2 years together	−0.0622	−0.0634	−0.0523
	(0.0389)	(0.0391)	(0.0424)
3 years together	−0.0720	−0.0757	−0.0621
	(0.0584)	(0.0599)	(0.0625)
4 years together	−0.0953	−0.1023	−0.0934
	(0.0737)	(0.0749)	(0.0781)
5 years together	−0.1536	−0.1720	−0.1478
	(0.0974)	(0.0991)	(0.0999)
6 years together	−0.2238	−0.2859	−0.2188
	(0.1160)	(0.1502)	(0.1277)
Experience profiles	1	2	2
Observations	139,962	139,750	139,749
Pairs	6,076	6,068	6,068

Notes: This table reports the second-step pair-tenure coefficients from the two-step estimator of Section 4.1, which regresses year-adjusted student test scores on indicators for years spent together as a pair, with pair fixed effects and controls for both teachers' total experience. The omitted category is the pair's first year together. Column (1) is the specification reported in the text, which absorbs a single experience profile common to all pairs. Columns (2) and (3) relax that restriction by absorbing a separate experience profile for teachers who formed their partnership early in their careers and for those who formed it later, where e_0 denotes a teacher's experience in the partnership's first observed year. Column (2) splits at $e_0 \leq 2$ years and column (3) at the median e_0 of eight years; each teacher is assigned to a group using their own e_0, so the general education and special education members of a pair may fall in different groups. Tenure levels above six years are absorbed by the fixed effects and are not reported. The sample includes all students in co-taught classrooms in Indiana from 2012–2019. Controls include student demographics and a cubic polynomial in the previous two years' test scores. Standard errors in parentheses are clustered at the pair level. No coefficient in any column is distinguishable from zero at conventional levels.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.3: Hazard Rate for Permanent Co-teaching Pair Separation



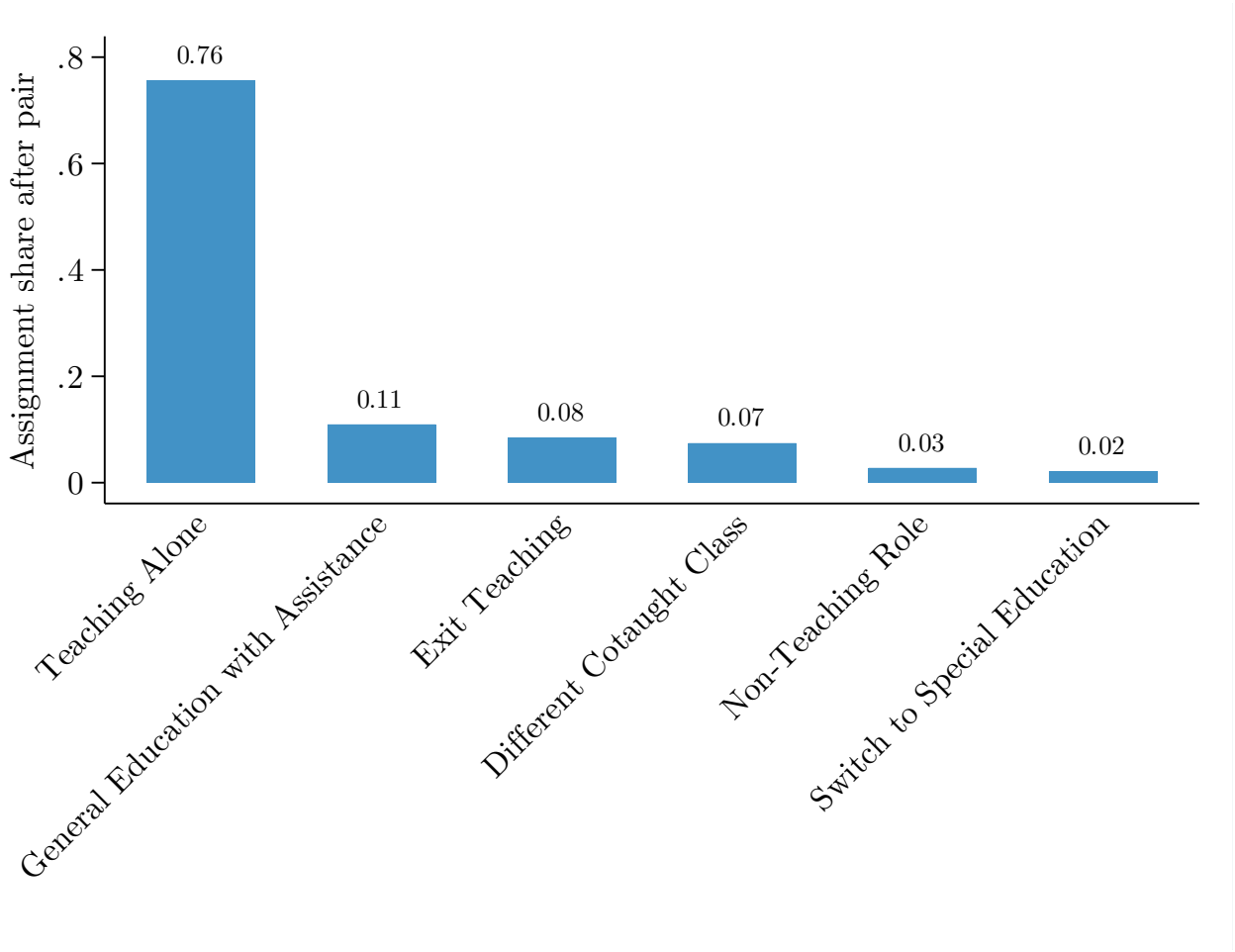
This figure shows the probability that a co-teaching pair permanently separates (i.e., never shares a classroom again in the sample period) rather than temporarily separating and later reuniting. Conditional on pair separation in a given year, the figure reports the share that never co-teach together again. The sample includes all co-teaching pairs in Indiana from 2012–2019.

(3.7) three years out. None is distinguishable from zero.[8]

Two features of the design further contain the concern. The stacked events are balanced within cohort-exit groups, so treated teachers and their controls span the same calendar window by construction, and differential attrition would have to operate within these matched groups. Beyond that, rather than sign the residual bias by assumption, I bound it.

[8]Retention has no informative pre-period, which is why the design conditions on survival rather than differencing against one. A teacher must be employed in year g to begin co-teaching in g, so treated teachers are retained in every earlier year by construction: their pre-period retention rate is 99.4 percent, against 87.0 percent for never-co-taught teachers. Any before-and-after comparison inherits that gap mechanically. The risk set therefore conditions both arms on employment in g and looks only forward. Conditioning instead on employment in $g-1$, as an earlier version of this design did, leaves treated teachers past the year-g exit hazard by construction while controls still face it, which inflates every horizon by roughly that hazard.

Figure A.4: What General Education Teachers Do After Co-Teaching

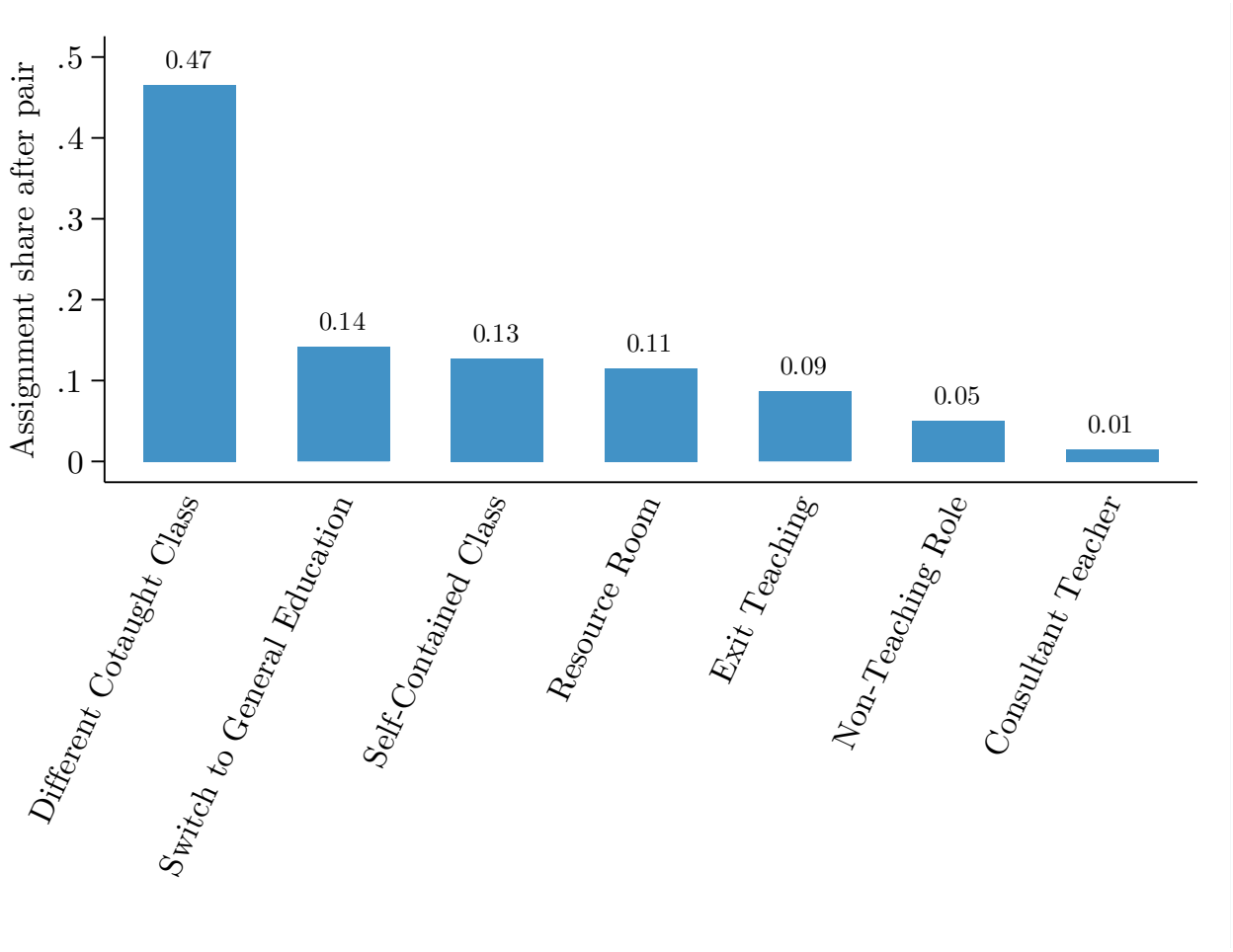


This figure shows the distribution of classroom assignments for general education teachers in the year immediately following the end of a co-teaching partnership. Categories include returning to solo instruction, entering a new co-teaching partnership, and leaving the sample (attrition). The sample includes all general education teachers in Indiana whose co-teaching partnership ended between 2012 and 2018.

The bound does not rest on that null. Because the retention estimates are imprecise enough to admit differences of either sign, I bound in both directions. If co-teaching retained teachers who would otherwise have left, the treated post-period sample carries extra marginal survivors, and the treated arm is the one to trim. If instead co-teaching pushed teachers out, the treated arm is depleted rather than padded and the control arm is over-represented, so the controls are what must be trimmed. I do both and report the union, which is the identified set when the direction of selection is unknown.

The trim shares are not chosen; they are read off the retention confidence intervals. At each horizon the treated arm is trimmed by the largest positive retention difference the

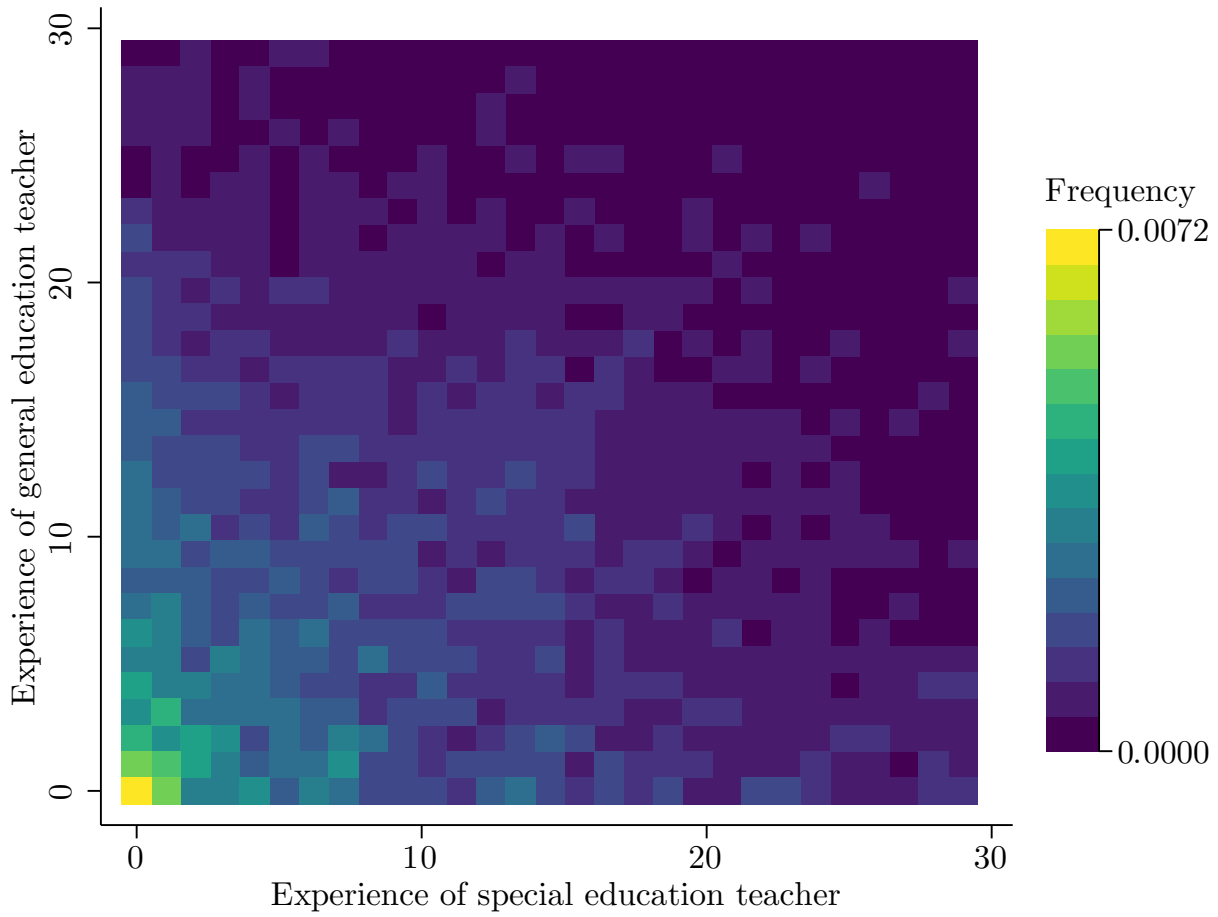
Figure A.5: What Special Education Teachers Do After Co-Teaching



This figure shows the distribution of classroom assignments for special education teachers in the year immediately following the end of a co-teaching partnership. Categories include returning to solo instruction, entering a new co-teaching partnership, and leaving the sample (attrition). The sample includes all special education teachers in Indiana whose co-teaching partnership ended between 2012 and 2018.

interval admits, and the control arm by the largest negative one, so each direction is stressed by exactly as much as the data allow. Those two numbers are very different. The upper edge of the retention interval is 1.6 percent at the two-year horizon, while the lower edge is 9.0 percent. Almost no positive retention difference is admissible, so there is little to trim from the treated arm; the bound is carried by the possibility that co-teaching pushed teachers out. Trimming the strongest treated teachers, or the weakest controls, is the worst case for a positive effect. I rank on achievement *net of prior scores*, so that "strong" and "weak" mean over- and under-performing baseline rather than teaching high-scoring students. The estimates hold: against an untrimmed 0.070σ, the union across both directions runs

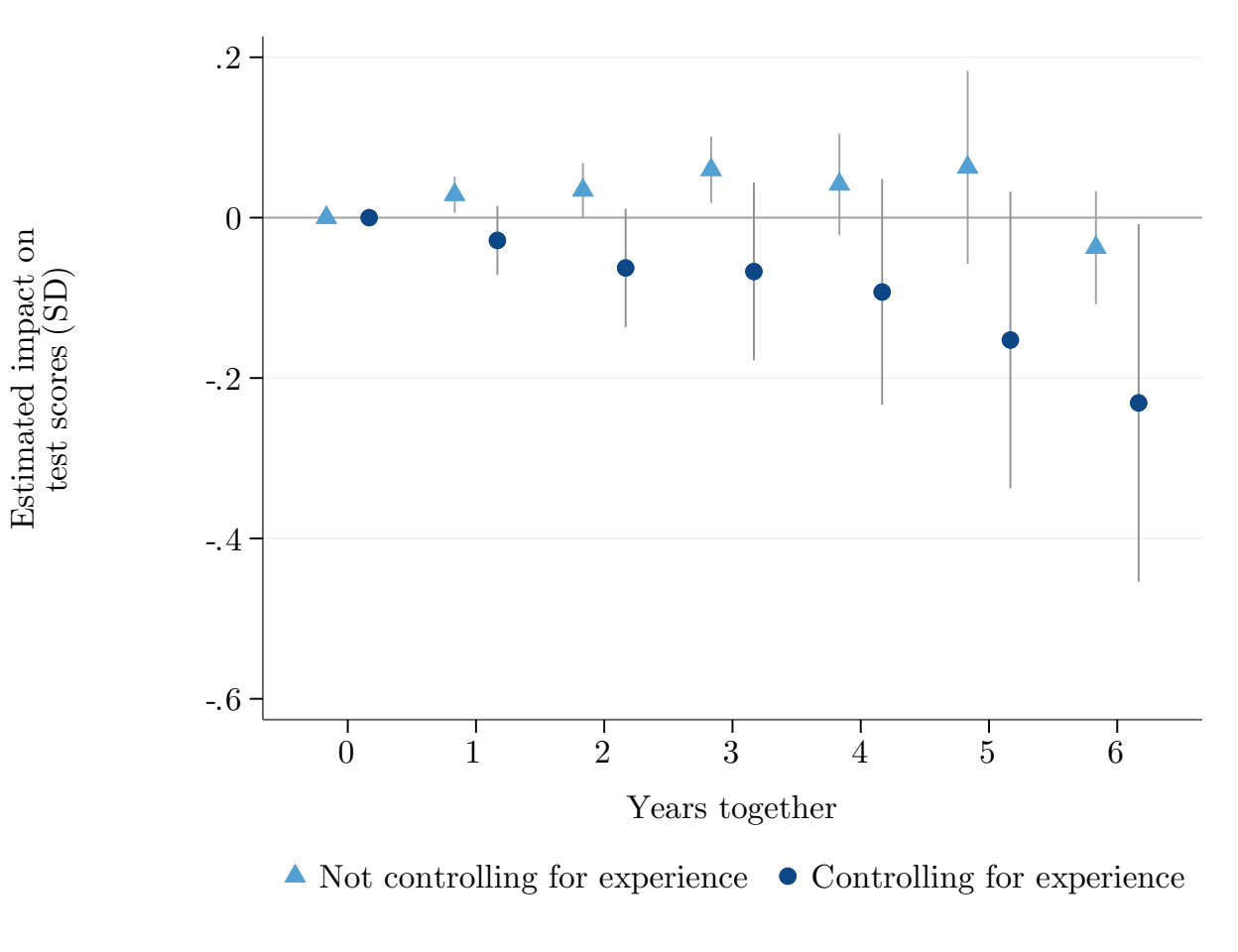
Figure A.6: Pairing Frequency by Experience



This heatmap shows the observed frequency of co-teaching pairs by the experience levels (in years) of the special education and general education teachers in the partnership. Each cell reports the share of all co-teaching pairs in Indiana (2012–2019) where the general education teacher has the row's experience level and the special education teacher has the column's experience level. Figure 3 shows the difference between these observed frequencies and those expected under random assignment.

$[0.06,\ 0.08]\sigma$ at two years, and the one-year bound stays positive throughout (Table B.1). Because the admissible treated-side trim is smaller than one teacher at the two-year horizon, that direction is close to non-binding, and the reported interval is effectively the worst case under the negative-selection branch.

Figure A.7: Return To Experience in a Pair (Estimated in One Step)

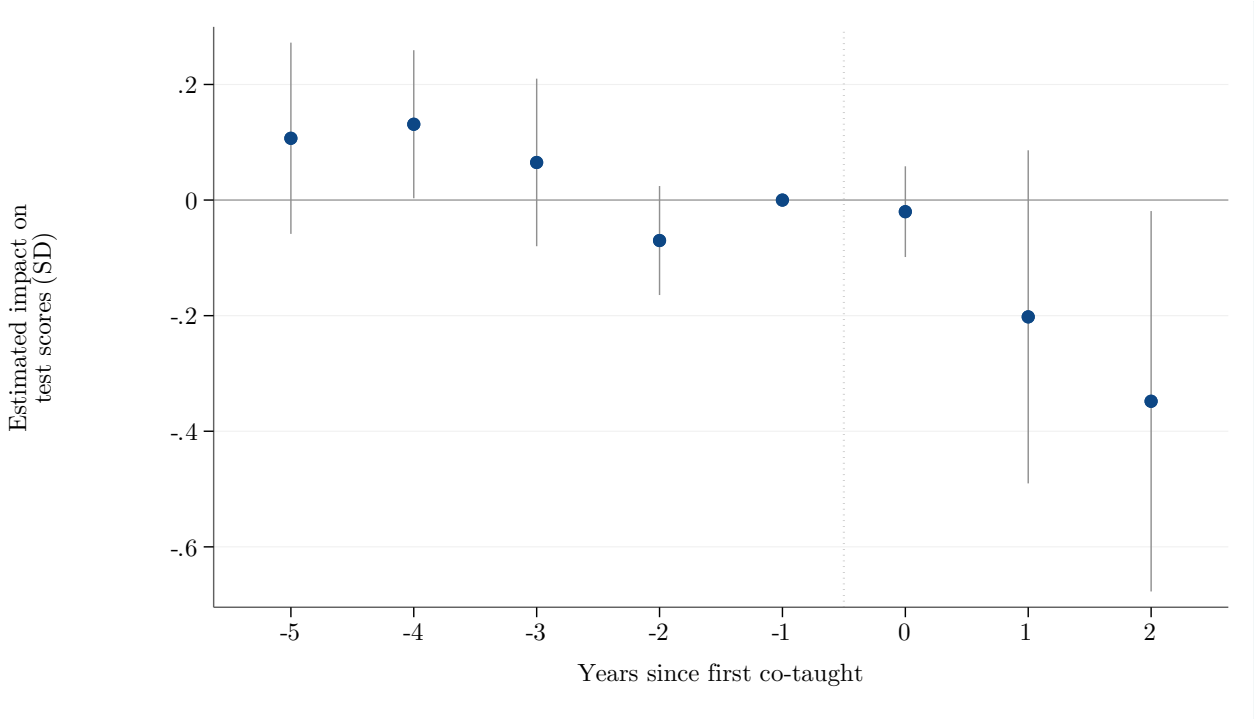


This figure presents the same analysis as Figure 4 but estimates the relationship between pair duration and student test scores in a single step rather than the two-step approach of Papay and Kraft (2015). The specification regresses student test scores on indicators for years spent in the current pair, student controls (demographics and a cubic polynomial of the previous two years' test scores), teacher fixed effects, and year fixed effects, with and without controls for general teaching experience. The sample includes all student-year observations in co-taught classrooms in Indiana from 2012–2019.

C Event Weighting, Estimator Choice, and Trend Sensitivity

This appendix reports the specification checks summarized in Section 6: how the stacked design weights events, whether an alternative estimator reproduces the results, how much violation of parallel trends the estimates can survive, and how the estimates build up across control layers.

Figure A.8: Difference in Student Test Scores By Time Since First Co-taught (Only Using More Than Three Years of Prior Experience)



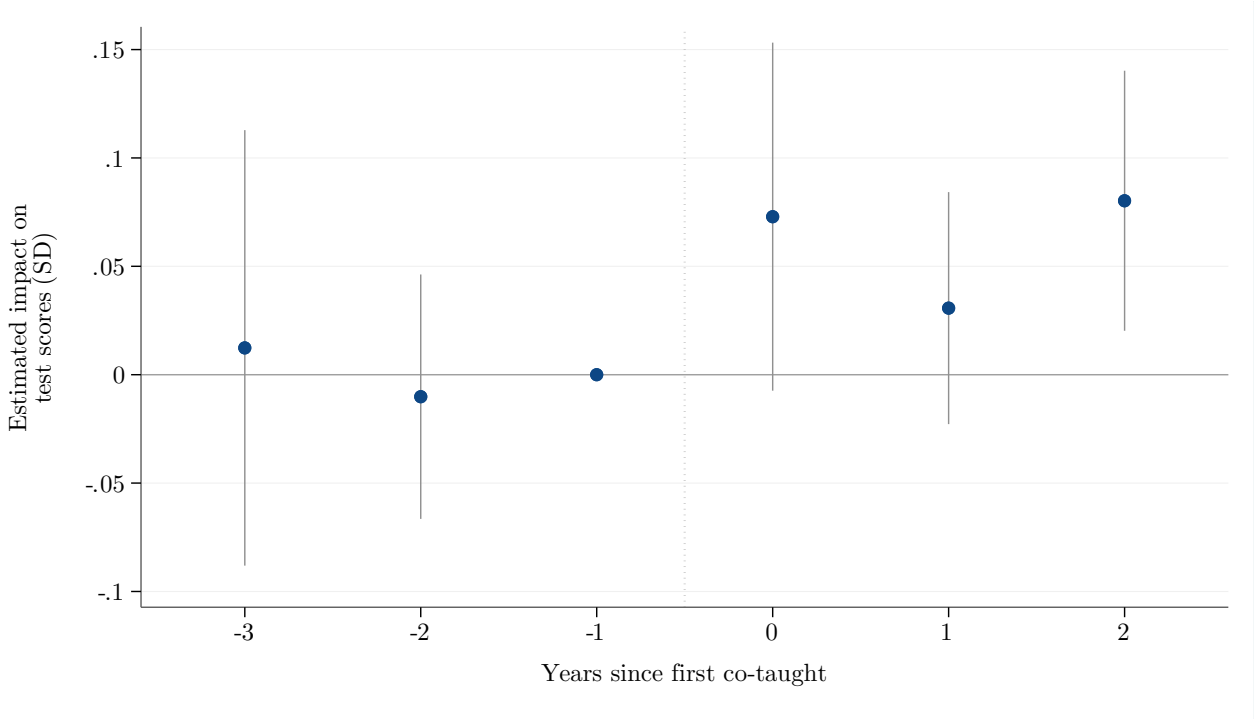
This figure replicates the main event study (Figure 5) but restricts the sample to teachers who had more than three years of solo teaching experience before first co-teaching. These teachers are excluded from the main analysis because they show different pre-trends, suggesting different selection into late-career co-teaching. The specification is otherwise identical: stacked difference-in-differences with student controls, event-specific teacher and year fixed effects, and standard errors clustered at the teacher level.

C.1 Event Weighting and Estimator Choice

The stacked difference-in-differences implicitly weights events, which could affect estimates if treatment effects vary. Table A.9 compares three approaches: implicit weights from the stacked methodology (0.037σ, s.e. 0.019), equal weights across events (0.041σ, s.e. 0.034), and weights proportional to observations per event (0.044σ, s.e. 0.022). The three estimates sit within a third of a standard error of one another, so the weighting scheme does not move the result. The implicit weights are not chosen to maximize precision. They are what the stacked design produces mechanically, and I report that specification as the main estimate for transparency.

I also re-estimate the design with the imputation estimator of Borusyak et al. (2024), which fits the untreated-outcome model on never- and not-yet-treated observations and imputes each treated observation's counterfactual. It is free of non-convex weighting by

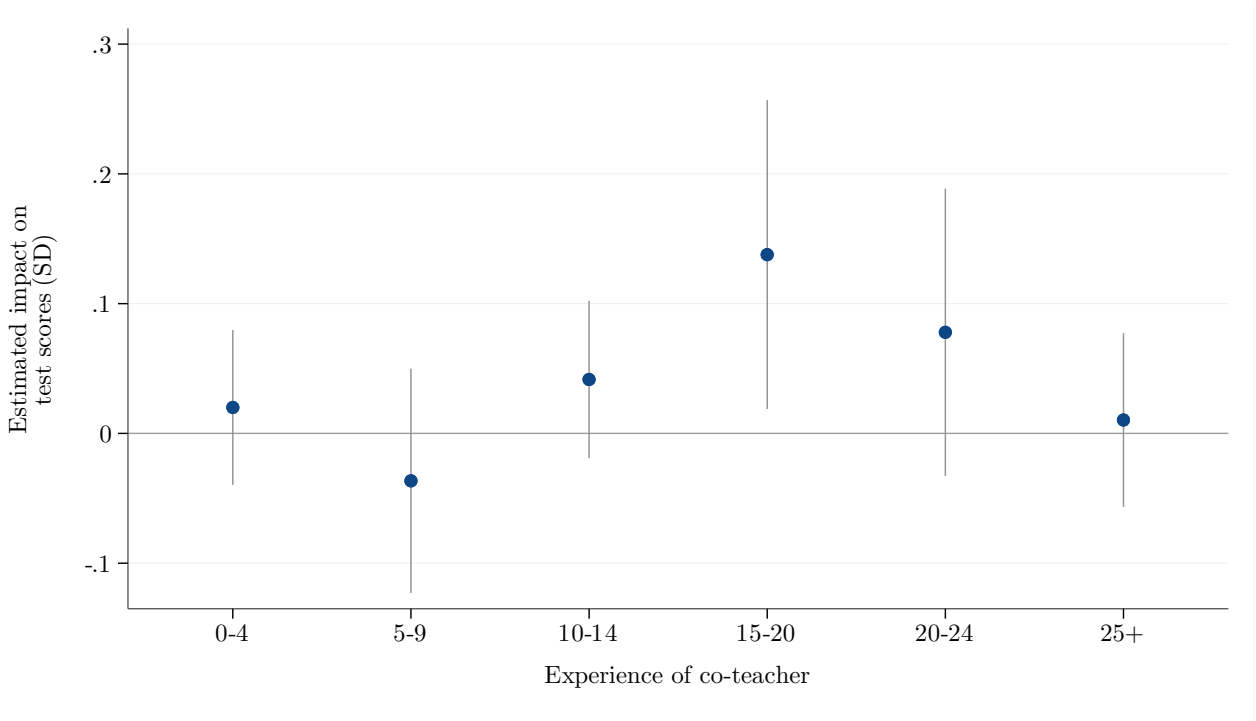
Figure A.9: Difference in Student Test Scores By Time Since First Co-taught (Controlling For Current Co-teaching)



This figure replicates the main event study (Figure 5) but adds indicator variables for whether the teacher is currently co-teaching to separate persistent post-co-teaching effects from contemporaneous collaboration benefits. The specification is otherwise identical: stacked difference-in-differences with student controls, event-specific teacher and year fixed effects, and standard errors clustered at the teacher level. Results are similar to the main specification, indicating that the estimated gains persist after co-teaching ends.

construction and accommodates the prior-score controls. I estimate it both for the average effect and separately by partner experience, since the experience gradient is the paper's central result. The imputation estimates reproduce the pattern (Appendix Figure A.19). The average post-period effects are positive (0.058σ, s.e. 0.038, in year 0) with pre-trend coefficients near zero. The partner-experience gradient is again clear: effects are indistinguishable from zero for partners with 0–7 years of experience, against 0.211σ (0.084), 0.463σ (0.182), and 0.195σ (0.081) in years 0 through 2 for 16+ partners. These estimates are larger and less precise than the stacked design's, as expected from a different estimand and control set, and the 16+ column shows one small significant pre-period coefficient (0.057σ two years out). I read the exercise as confirming that neither the average effect nor the gradient is an artifact of the stacked design's implicit weighting.

Figure A.10: Difference in Student Test Scores By Co-Teacher Experience



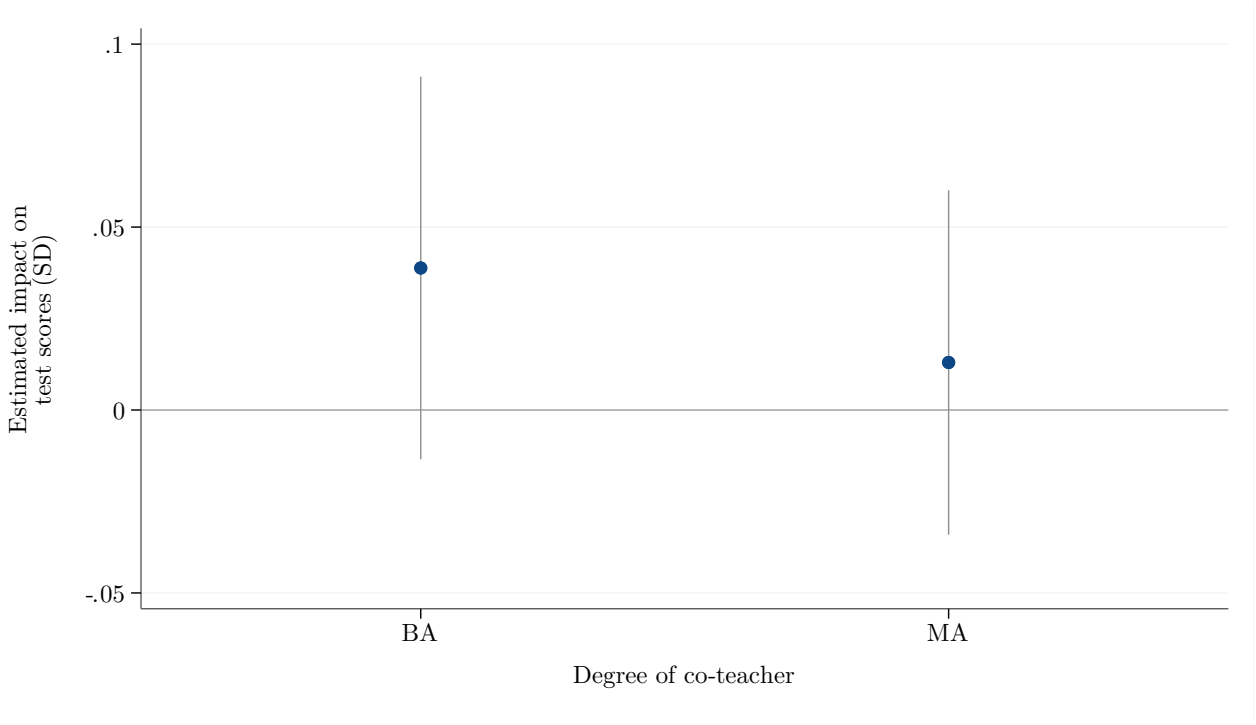
This figure estimates the effect of co-teaching on post-co-teaching student test scores by finer partner experience bins (5-year intervals). Each point represents a separate stacked difference-in-differences estimate comparing post-co-teaching to pre-co-teaching scores for teachers whose partner fell in the indicated experience range. The sample, controls, and fixed effects are identical to Figure 5. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

C.2 Sensitivity to Trend Violations and Controls

I next quantify how much hidden bias the results can survive, using the Rambachan and Roth (2023) relative-magnitudes bounds: robust confidence intervals that allow post-period violations of parallel trends up to $\bar{M}$ times the largest pre-period violation (Appendix Figure A.20). The two headline estimates fare differently. The year-0 *average* effect, whose conventional confidence interval already touches zero, has no margin against trend violations of this kind, since any admitted post-period violation pushes its interval through zero. The bounds stay centered on positive effects, and the estimate itself is stable across weighting schemes, control sets, and estimators. This is a limit on precision under the relative-magnitudes restriction for a borderline $t \approx 2$ estimate, not evidence that the average is fragile to specification.

The year-0 *16+ partner* effect is sturdier. Its robust interval excludes zero for violations

Figure A.11: Difference in Student Test Scores By Co-teacher Degree



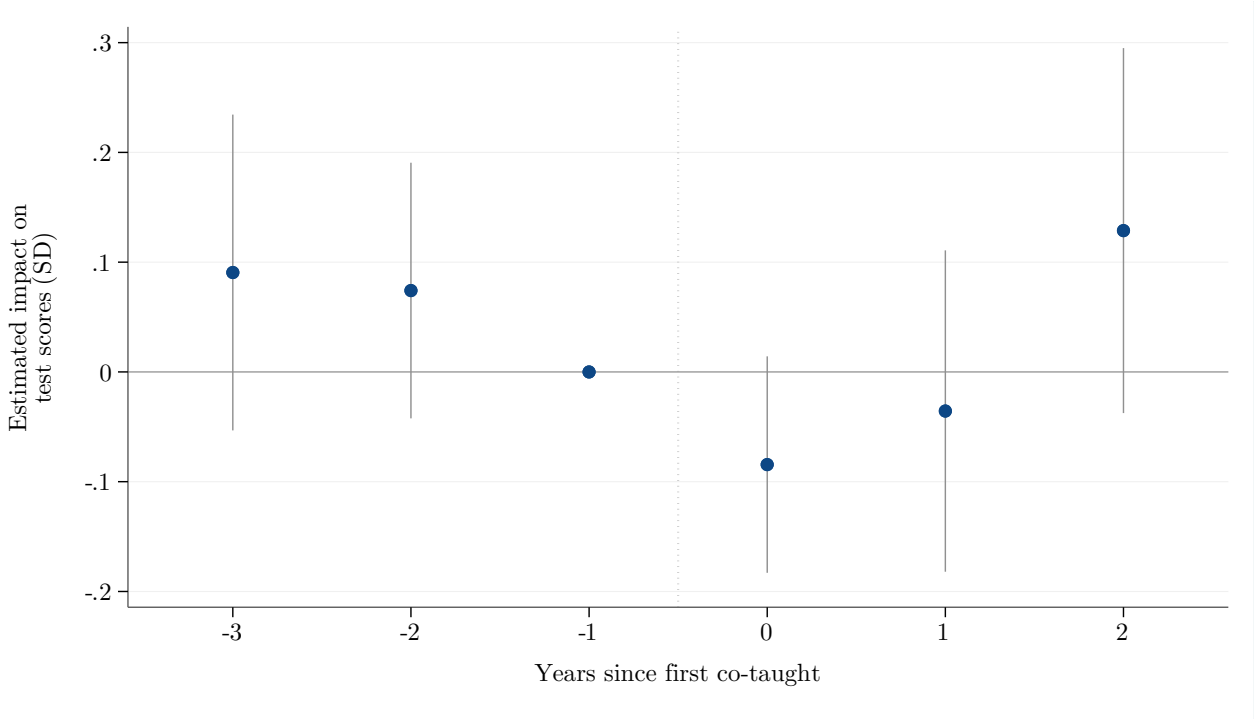
This figure estimates whether the co-teaching partner's educational attainment (bachelor's, master's, or doctorate) predicts the magnitude of developmental benefits. Each point represents a separate stacked difference-in-differences estimate for teachers grouped by their co-teacher's highest degree. The specification follows Equation (3) with a single post-co-teaching indicator. Controls and fixed effects are the full set given in Equation (3). Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

up to $\bar{M} = 0.75$. That standard is more demanding than the ratio suggests, because the pre-period anchor is small in absolute terms: the largest consecutive pre-period violation is roughly 0.02σ, so the bound tolerates post-period violations of only about 0.015σ per year. Pooling the three post-period effects is mechanically more conservative under relative magnitudes, because admissible violations compound with horizon. The pooled 16+ interval excludes zero for $\bar{M} \leq 0.25$.

The sensitivity analysis concentrates the paper's evidentiary weight where the paper already places it, on the partner-experience gradient and on the near-term effects within it. It also disciplines the policy simulation of Section 7. The projected returns inherit the estimates' sensitivity to trend violations, which is why the simulation's conservative scenarios, built on effects well below the point estimates, carry as much interpretive weight as the baseline.

The estimate is also not an artifact of which controls enter the model. Building the

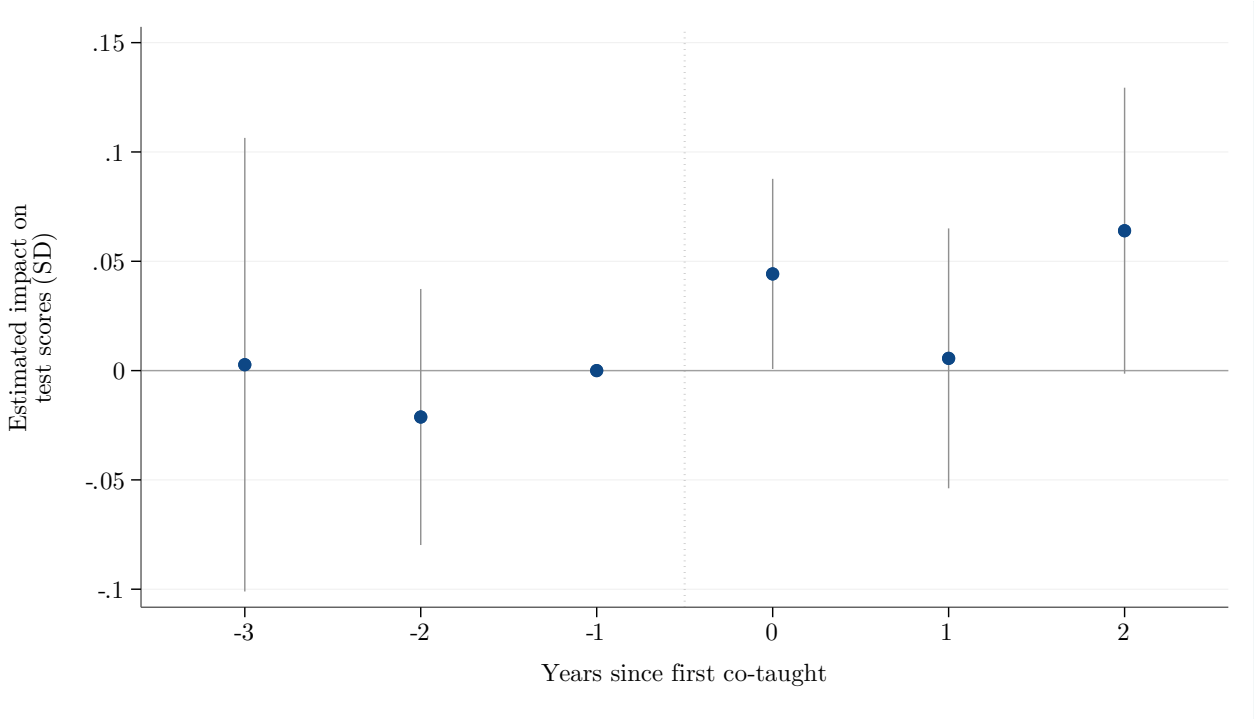
Figure A.12: Difference in Student Test Scores By Time Since First Co-taught (No Lagged Test Scores)



This figure replicates the main event study (Figure 5) but omits all student prior test score controls from the specification. Without conditioning on prior achievement, the positive effects disappear, consistent with downward bias from changes in student composition (co-teachers receive more disadvantaged students). The specification retains student demographics, event-specific teacher and year fixed effects, and standard errors clustered at the teacher level.

specification up sequentially (Appendix Table A.8) shows one layer doing the work. With event fixed effects alone the post-period estimates are noisy and signless, the biased no-controls benchmark of Figure A.12. Adding the prior-score cubics produces the positive post-period pattern. The remaining layers, demographic fixed effects and classroom composition, move the estimates by a few thousandths of a standard deviation. The same holds within the 16+ partner subsample, where the year-2 estimate goes from 0.215σ with prior scores alone to 0.156σ in the full specification, significant throughout. Nor does imposing common control coefficients across sub-experiments matter. Allowing the prior-score controls to differ by event, roughly 850 additional coefficients, moves the year-0 estimate from 0.039σ to 0.042σ and the year-2 estimate from 0.070σ to 0.063σ, with pre-period coefficients unchanged. The never-co-taught control teachers overlap heavily

Figure A.13: Difference in Student Test Scores By Time Since First Co-taught (One Year of Lagged Test Scores)



This figure replicates the main event study (Figure 5) but uses only one year of prior student test scores (a cubic polynomial of $t-1$ scores) rather than the two years used in the main specification ($t-1$ and $t-2$ scores). Results are similar, indicating that the choice of lag structure does not drive the findings. The specification retains student demographics, event-specific teacher and year fixed effects, and standard errors clustered at the teacher level.

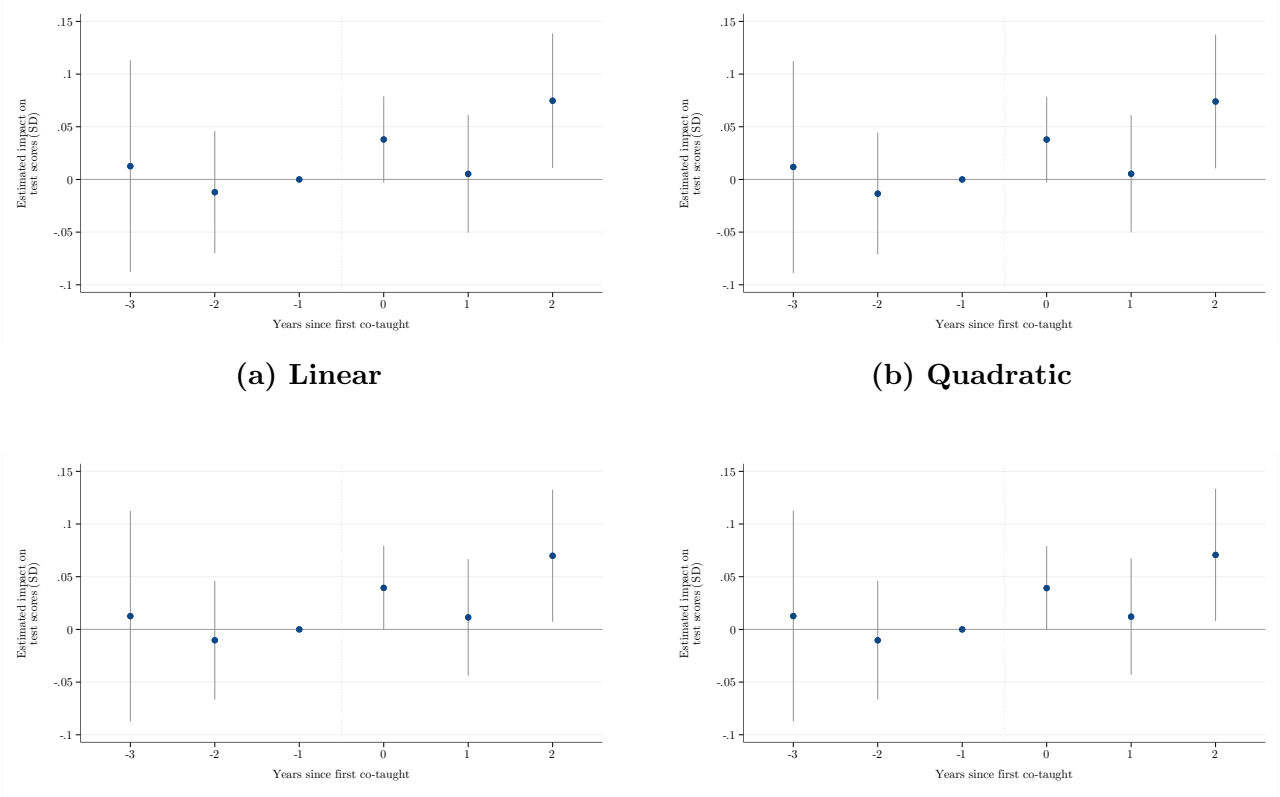
across events, which disciplines the event-specific coefficients toward their pooled values.[9]

D Estimating Equations and Identifying Assumptions

This appendix collects, in one place, the estimating equation and identifying assumption for every empirical exercise in the paper. The two specifications already written out in the main text, the two-step pair-tenure decomposition (Section 4.1) and the stacked event study of Equation (3), are not repeated; the entries below cover the remaining analyses in order of appearance. Purely descriptive comparisons of means (Tables A.1, A.4, and the summary statistics) involve no estimating equation and are omitted.

[9] A joint test of the interactions is not feasible with teacher-clustered inference; the interaction coefficients outnumber the cluster degrees of freedom.

Figure A.14: Main Event Study Under Alternative Prior-Score Polynomial Orders



(a) Linear **(b) Quadratic**

(c) Cubic (main specification) **(d) Quartic**

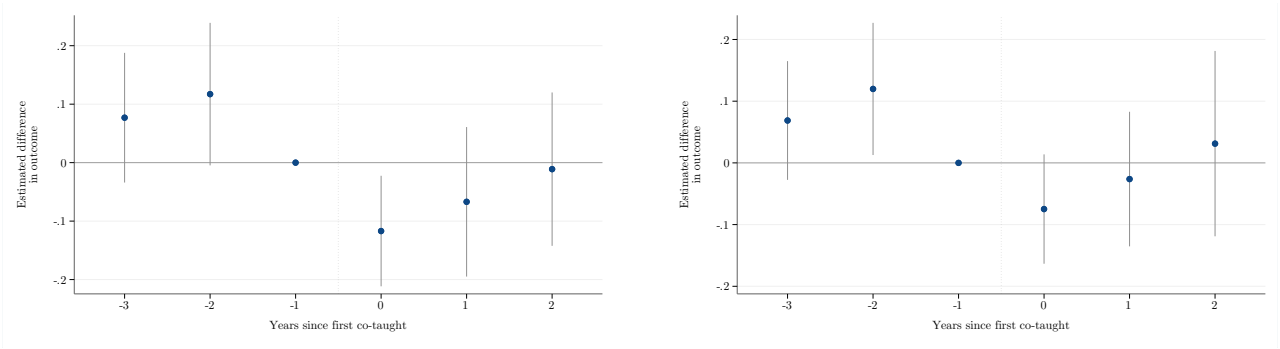
This figure replicates the main event study (Figure 5) varying the polynomial order of the prior test score controls from linear through quartic. The polynomial applies to all four lagged-score variables ($t-1$ and $t-2$ scores in math and ELA); panel (c) reproduces the main specification's cubic. All other controls are identical: student demographics, classroom composition, event-specific teacher and year fixed effects. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

D.1 Co-Teaching Response to Special-Education Enrollment (Figure A.1)

Let c index a school-grade cell and y a year, let CoTaught_{cy} equal one if the cell has a co-taught class, and write N_{cy} for the cell's special-education count, $\Delta N_{cy} = N_{cy} - N_{c,y-1}$ for its year-over-year change, and $L_{cy} = N_{c,y-1}$ for the starting level. The specification interacts the change with a quadratic in the starting level,

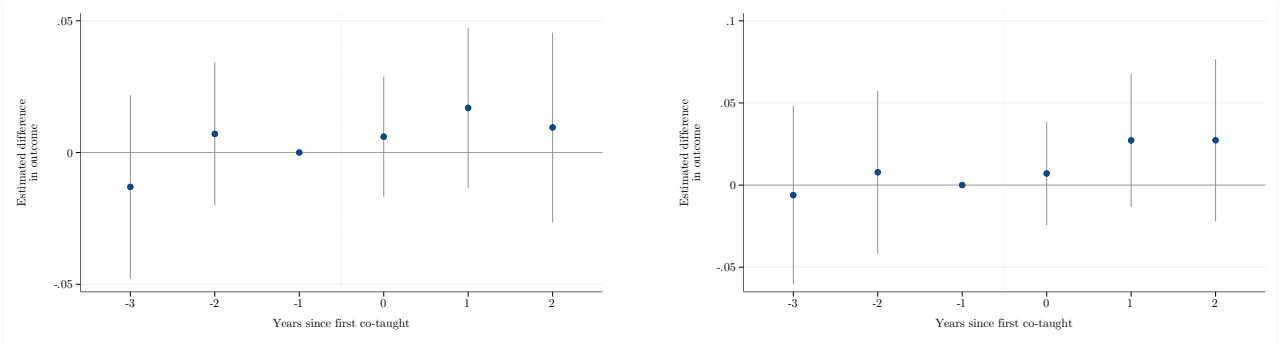
$$\text{CoTaught}_{cy} = \left(\beta_1 + \beta_2 L_{cy} + \beta_3 L_{cy}^2\right) \Delta N_{cy} + f(L_{cy}) + \phi_c + \tau_y + \varepsilon_{cy}, \tag{4}$$

Figure A.15: Student Composition Around Co-Teaching Entry



(a) Lagged Math Score

(b) Lagged ELA Score



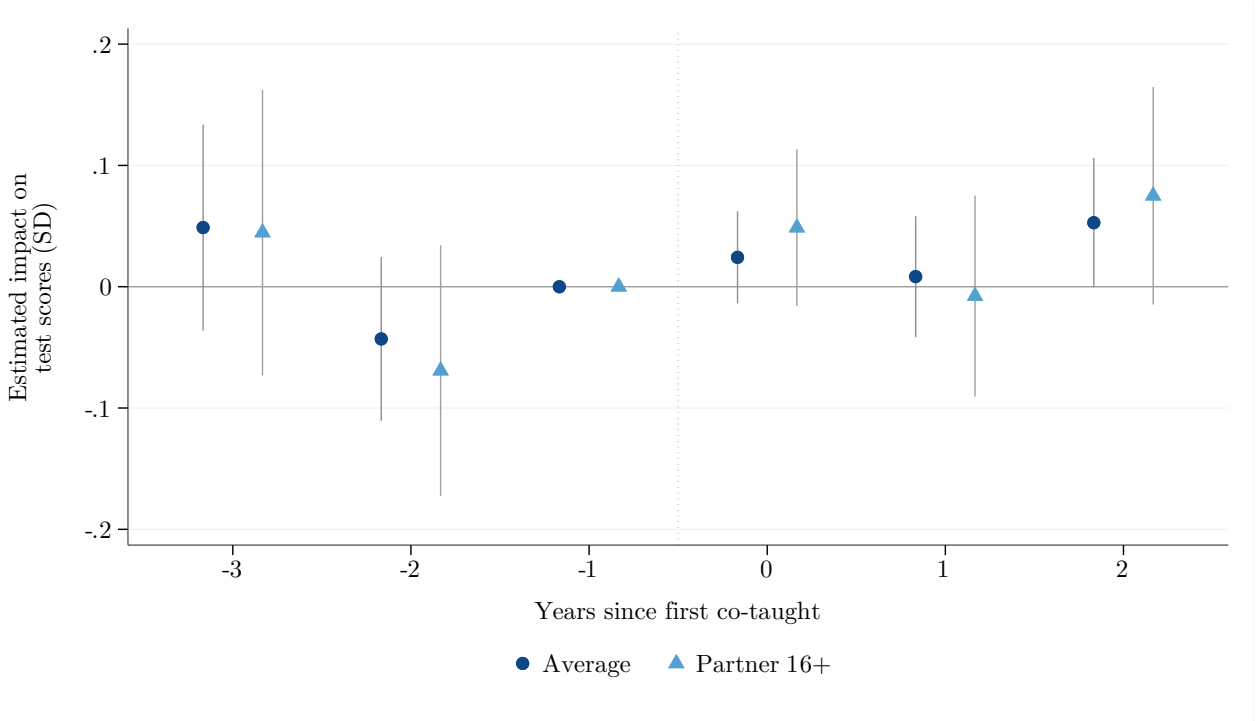
(c) Ever English Learner

(d) Student of Color

This figure estimates the main event study with student characteristics as outcomes rather than test scores, checking whether classroom composition shifts discretely when a teacher begins co-teaching. Each panel regresses the indicated characteristic on event-time indicators with event-specific teacher and year fixed effects (no student controls, since the characteristics are the outcomes). Flat profiles indicate that the student sorting documented in Table 2 is a level difference rather than a discrete change at co-teaching entry. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

where ϕ_c and τ_y are school-grade and year fixed effects, $f(L_{cy})$ collects the lower-order terms in the starting level, and standard errors are clustered at the school. Figure A.1 plots the marginal effect $\partial \text{CoTaught}_{cy}/\partial \Delta N_{cy} = \beta_1 + \beta_2 L_{cy} + \beta_3 L_{cy}^2$ across L_{cy}. The exercise is descriptive. It does not assert that enrollment is exogenous to teacher quality; the identifying content is only that, within a cell and net of common year shocks, an additional special-education student raises the probability of co-teaching, by an amount that declines in the starting count, which is the extensive-margin pattern the section invokes.

Figure A.16: Main Event Study With Student Fixed Effects



This figure estimates the main event study replacing the prior-score controls with student fixed effects. Each event is augmented with its students' test-score observations from non-event teachers' classrooms in grades 3–8; these anchor observations carry their own event-time category and event-specific teacher effects, so they identify the student fixed effects without contributing to the treated-control contrast. Post-co-teaching years of anchor teachers who themselves co-taught are absorbed by event-specific indicators. Estimates are within teacher, year, and student, with classroom composition controls on event observations. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

D.2 Student-Composition Difference-in-Differences (Table 2)

For teacher t's classroom in year y, let S_{ty} be the share of students with a given characteristic. With Ever_t marking teachers who ever co-teach and Post_{ty} marking years at or after first co-teaching,

$$S_{ty} = \beta_0 + \beta_1 \text{Ever}_t + \beta_2 \text{Post}_{ty} + \beta_3 \left(\text{Ever}_t \times \text{Post}_{ty}\right) + \nu_{ty}, \tag{5}$$

with standard errors clustered at the teacher. The coefficient of interest is β_3, the within-teacher change in composition coinciding with co-teaching entry. The regression is descriptive: it documents that observable student inputs shift at entry, which is why the achievement specifications condition on prior scores. Because post years mix co-taught and returned-to-solo classrooms, β_3 understates the contrast between co-taught and solo rooms.

Table A.6: Student Fixed Effects vs. Lagged-Score Controls, Unstacked Panel

	(1)	(2)	(3)	(4)
	Student FE	Lagged scores	Student FE	Lagged scores
Partner experience	All	All	16+	16+
Post-co-teaching (analysis sample)	0.047**	0.050**	0.081**	0.109***
	(0.015)	(0.019)	(0.025)	(0.032)
Post-co-teaching (other co-teachers)	0.000	−0.003	0.001	−0.003
	(0.006)	(0.008)	(0.006)	(0.008)
Currently co-taught classroom	−0.003	−0.002	−0.004	−0.000
	(0.004)	(0.005)	(0.004)	(0.005)
Observations	2,902,818	2,961,035	2,869,326	2,932,418

Notes: This table compares the two control strategies on the unstacked panel of all teachers' student-year observations in tested grades (Math and ELA, 2013–2019). The outcome is the standardized test score. "Post-co-teaching (analysis sample)" indicates years from first co-teaching onward for teachers meeting the event-study criteria (known first year, first co-taught 1–3 years after entry); columns (3)–(4) restrict these teachers to those whose partner had 16+ years of experience. "Post-co-teaching (other co-teachers)" indicates post-co-teaching years of all other ever-co-teaching teachers. Odd columns absorb student, teacher, year, and grade-by-subject fixed effects with classroom composition controls; even columns replace the student fixed effects with the main specification's lagged-score controls. Identical samples up to singleton drops, so differences across columns reflect the control strategy alone. Standard errors in parentheses are clustered at the teacher level.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The dynamic version (Figure A.15) is Equation (3) estimated with a student characteristic as the outcome and no student controls, checking that the shift is a level difference rather than a discrete jump at entry.

D.3 Test of Strategic Pairing (Section 3)

The benchmark is built on dyadic pairs: each co-teaching pair-year is oriented by role into one general-education and one special-education member using the majority of its class-level setting codes, and the few pairs coded the same role on both sides are excluded, having no cell on a general-by-special grid. Let O_{jk} be the observed count of co-teaching pairs with general-education experience j and special-education experience k, and E_{jk} the count expected under random within-year matching, where E_{jk} is the year-weighted product of that year's marginal experience distributions summed over years. The Pearson statistic $\sum_{jk}(O_{jk} - E_{jk})^2/E_{jk}$ over cells with $E_{jk} \geq 5$ tests the null of random pairing. Holding the

Table A.7: Experienced-Partner Effect Across Balanced-Subset Event-Study Specifications

		Required		16+ partner effect, event time (σ)					
gap=0	k=0	Pre	Post	-3	-2	0	$+1$	$+2$	N
No	No	—	$+2$					$+0.181^{***}$	45
No	No	—	$+1$				$+0.087^{*}$		76
No	No	—	$+1, +2$				$+0.143^{**}$	$+0.173^{***}$	38
No	No	-2	$+2$		$+0.133^{*}$			$+0.220^{**}$	10
No	No	-2	$+1$		$+0.049$		$+0.068$		20
No	No	-2	$+1, +2$		$+0.128^{*}$		$+0.046$	$+0.220^{**}$	9
No	No	$-3, -2$	$+2$	$+0.159$	$+0.100$			$+0.169^{*}$	4
No	No	$-3, -2$	$+1$	$+0.123$	$+0.083$		$+0.111$		8
No	No	$-3, -2$	$+1, +2$	$+0.174$	$+0.106$		$+0.042$	$+0.198^{*}$	3
No	Yes	—	$+2$			$+0.208^{***}$		$+0.175^{***}$	38
No	Yes	—	$+1$			$+0.111^{**}$	$+0.070$		69
No	Yes	—	$+1, +2$			$+0.219^{***}$	$+0.125^{**}$	$+0.155^{***}$	34
No	Yes	-2	$+2$		$+0.136^{*}$	$+0.182^{*}$		$+0.205^{**}$	10
No	Yes	-2	$+1$		$+0.036$	$+0.094$	$+0.066$		20
No	Yes	-2	$+1, +2$		$+0.141^{*}$	$+0.173^{*}$	$+0.054$	$+0.210^{**}$	9
No	Yes	$-3, -2$	$+2$	$+0.179$	$+0.111$	$+0.166$		$+0.186^{*}$	4
No	Yes	$-3, -2$	$+1$	$+0.106$	$+0.060$	$+0.182$	$+0.114$		8
No	Yes	$-3, -2$	$+1, +2$	$+0.177$	$+0.109$	$+0.154$	$+0.044$	$+0.197^{*}$	3
Yes	No	—	$+2$					$+0.179^{***}$	105
Yes	No	—	$+1$				$+0.087^{*}$		160
Yes	No	—	$+1, +2$				$+0.170^{***}$	$+0.144^{***}$	88
Yes	Yes	—	$+2$			$+0.193^{***}$		$+0.184^{***}$	87
Yes	Yes	—	$+1$			$+0.108^{**}$	$+0.068$		140
Yes	Yes	—	$+1, +2$			$+0.160^{***}$	$+0.169^{***}$	$+0.165^{***}$	76

Notes: Each row is a separate stacked event study of student test scores on years since first co-teaching, restricted to teachers paired with a 16+ year partner (or never co-taught), on the known-first-year sample. The panel makes horizons beyond year two unobservable, so each specification instead imposes a balanced/rectangular panel on a chosen *subset* of the remaining event times: the "Required" columns list which pre-periods and post-horizons a teacher must be observed at to enter, and whether year zero (the co-teaching year) is included; event times outside the required subset are ignored, so looser subsets retain more teachers. "gap=0" indicates whether the year-zero-cotaught cohort (co-teach in the imputed entry year, hence no pre-period) is included; because that cohort has no pre-period it is mechanically absent whenever a pre-period is required, so exclude and include coincide there and only the no-pre-period rows report both. N is the number of treated teachers with a 16+ partner. Because the balance requirements here differ from those defining the main event-study sample, N is not bounded by the 141 treated teachers of Table A.2: specifications that require only a single post-horizon and admit the year-zero-cotaught cohort retain teachers whom the main sample's balanced window excludes. Of the 18 pre-period placebo coefficients across these specifications, 4 are individually significant at the five percent level but none survives a Benjamini–Hochberg correction at a five percent false-discovery rate. Standard errors are clustered at the teacher level.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.8: Sequential Control Build-Up

	(1) Event FEs	(2) + Prior scores	(3) + Demographics	(4) + Class comp
Panel A: All partners				
Year 0	−0.084	0.006	0.033	0.039
	(0.050)	(0.025)	(0.021)	(0.020)
Year 1	−0.036	0.021	0.013	0.011
	(0.075)	(0.038)	(0.029)	(0.028)
Year 2	0.129	0.106**	0.077*	0.070*
	(0.085)	(0.041)	(0.033)	(0.032)
Observations	171,824	171,824	156,025	156,025
Panel B: Partner experience 16+				
Year 0	0.090	0.078*	0.115**	0.118**
	(0.066)	(0.039)	(0.036)	(0.037)
Year 1	−0.019	0.039	0.059	0.062
	(0.150)	(0.066)	(0.048)	(0.048)
Year 2	0.391*	0.215**	0.162**	0.156**
	(0.173)	(0.074)	(0.051)	(0.048)
Observations	106,322	106,322	96,646	96,646

Notes: This table adds controls to the stacked event study one layer at a time. Column (1) includes only the event-specific teacher and year fixed effects (with class-size and grade-by-subject effects); column (2) adds the cubic polynomials in the previous two years' math and ELA scores; column (3) adds the demographic fixed effects (special education and ELL history, status-transition paths, ethnicity, gender, and the school-by-free-lunch-status effect); column (4) adds classroom composition shares, reproducing the main specification. Rows report the post-period event-time coefficients (years 0–2 relative to year −1). Panel B restricts treated teachers to those whose partner had 16+ years of experience. Observation counts fall in columns (3)–(4) because the demographic fixed effects require nonmissing demographic data. Standard errors in parentheses are clustered at the teacher level.

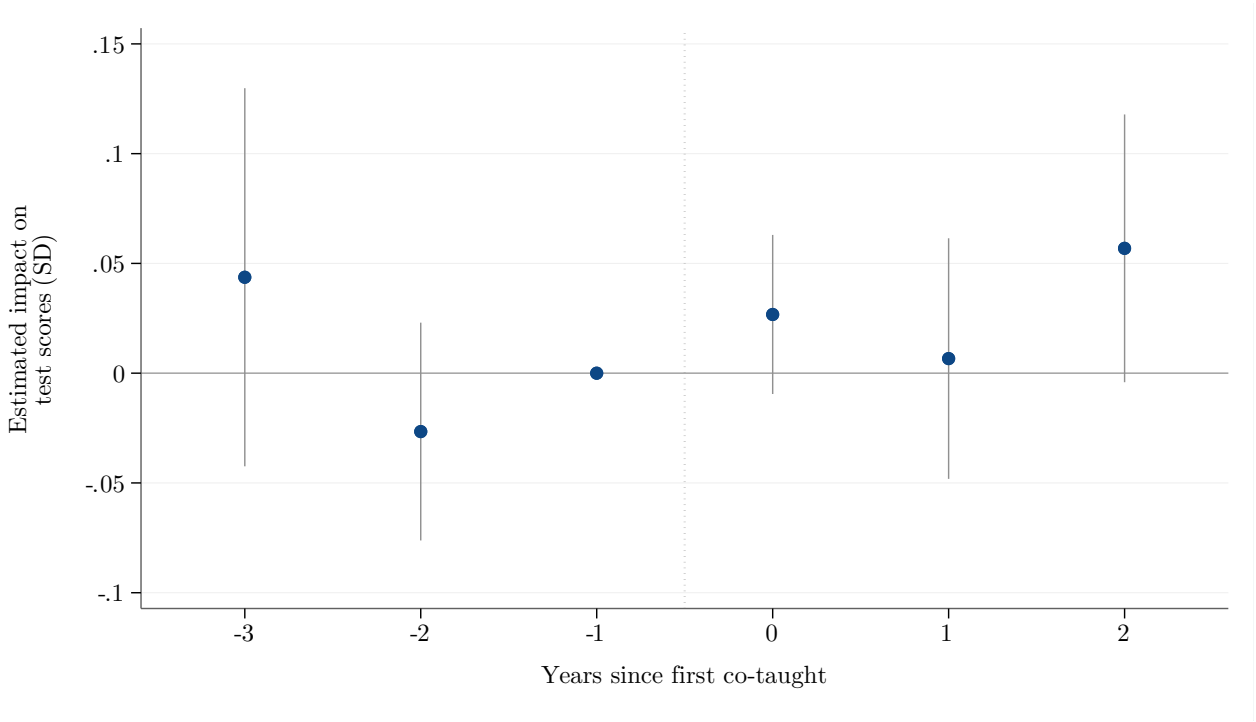
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.9: Difference-in-Differences Estimates Under Alternative Event Weighting

	Student Test Scores		
	Implicit Weights	Even Weights	Observation Weights
Estimated Difference	0.031	0.017	0.055

Notes: This table compares the main difference-in-differences estimate under three weighting schemes. "Implicit Weights" uses the weights arising naturally from the stacked estimation, "Even Weights" reweights so each event contributes equally, and "Observation Weights" weights events proportional to their number of student-year observations. All specifications use the event study sample (general education teachers with 1–3 years of prior solo experience), control for student demographics and a cubic polynomial of the previous two years' test scores, and include event-specific teacher and year fixed effects. Standard errors are clustered at the teacher level.

Figure A.17: Difference in Student Test Scores By Time Since First Co-taught (One or More Years of Prior Experience)



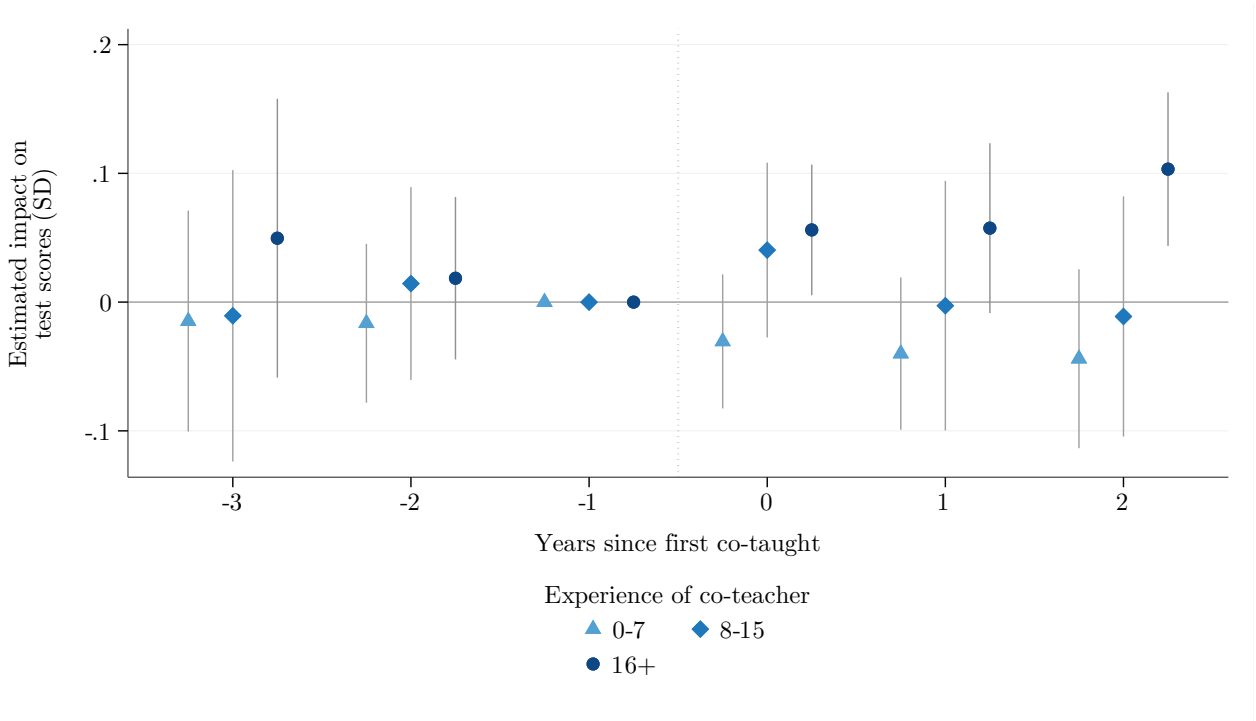
This figure replicates the main event study (Figure 5) but expands the sample to include all general education teachers with at least one year of prior solo experience before co-teaching, rather than restricting to 1–3 years. The broader sample slightly attenuates estimated effects but preserves the qualitative pattern. The specification is otherwise identical: stacked difference-in-differences with student controls, event-specific teacher and year fixed effects, and standard errors clustered at the teacher level.

Table B.1: Worst-Case Attrition Bounds on the Post-Co-Teaching Effect

Horizon	Point estimate	Trim shares (treated / control)	Worst-case bound
Year 1 ($k = 1$)	0.011	1.3% / 5.7%	[0.007, 0.017]
Year 2 ($k = 2$)	0.070	1.6% / 9.0%	[0.062, 0.077]

Notes: Trimming bounds on the stacked event-study coefficients at post horizons $k = 1$ and $k = 2$. Because the risk-set estimates admit retention differences of either sign, bounds are computed in both directions and the union is reported. Trimming the treated arm covers the case where co-teaching retains marginal teachers; trimming the controls in the same event-year cells covers the case where it pushes them out. The share trimmed in each direction is the corresponding edge of that horizon's 95 percent confidence interval on the retention difference, so each branch is stressed by the largest difference the data admit in that direction. For the lower bound the strongest treated teachers, or the weakest controls, are dropped; for the upper bound the reverse. Trimmed observations are removed only at the horizon being bounded, and the event study is re-estimated on the remaining sample. Outcomes are residualized on the full lagged-score and composition controls before ranking.

Figure A.18: Partner-Experience Gradient on the Imputed-Entry Sample (Cohorts to 2006, Any Career Stage)



This figure re-estimates the partner-experience event study on a broad sample that recovers the teachers the known-first-year restriction drops. Each teacher's entry year is imputed from the administrative experience field as the minimum over observed years of (school year − years of experience). The sample includes imputed-entry cohorts back to 2006 and imposes no restriction on how long after entry a teacher first co-teaches, so the design includes teachers who co-teach as veterans, the population the panel's 2012 start otherwise excludes. This yields 393 treated teachers (roughly 2.8 times the main sample), 246 of them teachers with no known first year. The experienced-partner (16+) gradient persists on this sample: 0.056σ (s.e. 0.026) in year 0 and 0.103σ (0.030) in year 2, both significant, while the pooled average attenuates to 0.019σ and partners with 0–7 years of experience produce no gains. Because the co-taught subject is not identified on the unrestricted file, the sample uses all co-teaching pairs rather than same-subject pairs as in the primary specification. The plotted window is truncated at $k = -3$ to match the main specification. Because this sample imposes no restriction on experience at entry, earlier event times are estimable here, but only for teachers with four or more years of prior experience, whom the primary sample excludes; those cells are correspondingly thin and carry the entry selection documented in Section 6.2. Specification, controls, and fixed effects otherwise match the main event study; standard errors are clustered at the teacher level.

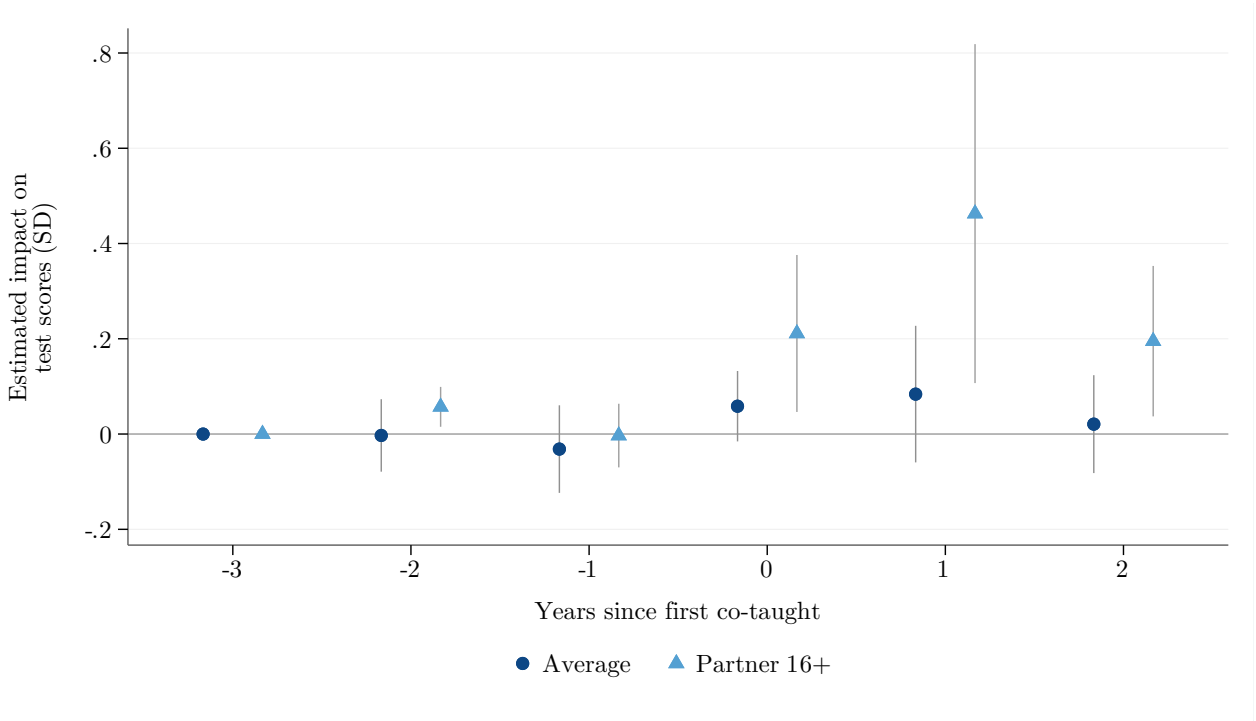
year-specific pool fixed means a rejection reflects assortativity in who pairs with whom, not the composition of the available teachers. Identification of the main results does not require this null to hold; the test shows only that what assortativity exists runs against, not toward, strategic mentorship.

D.4 Returns to Pair Duration (Table 3)

For student i in pair p's classroom in year y,

$$\pi_{ipy} = \theta \, \text{Duration}_p + \lambda_y + X'_{ipy}\beta + \left(\mu^g_{g(p)} + \mu^s_{s(p)}\right) + \varepsilon_{ipy}, \tag{6}$$

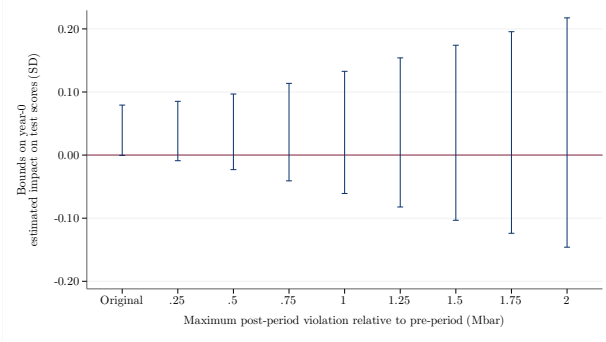
Figure A.19: Imputation Estimates of the Main Event Study



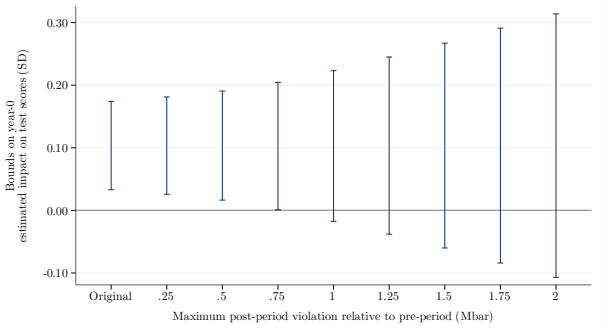
This figure presents Borusyak et al. (2024) imputation estimates of the effect of co-teaching on student test scores, for the average effect and for teachers whose partner had 16+ years of experience. The estimator fits the untreated-outcome model on never- and not-yet-treated observations, including the prior-score controls and teacher, year, and grade-by-subject fixed effects, and imputes each treated observation's counterfactual; it is free of the non-convex weighting concern that applies to stacked designs. The sample mirrors the event-study population (known-first-year teachers who first co-taught one to three years after entry, or never co-taught; Math and ELA). Pre-period coefficients are placebo estimates. Error bars represent 95% confidence intervals with standard errors clustered at the teacher level.

estimated without (Panel A) and with (Panel B) separate fixed effects for the pair's general-education teacher $g(p)$ and special-education teacher $s(p)$; X holds student demographics and a cubic in the previous two years' scores, and standard errors cluster at the pair. Completed duration Duration_p is fixed within a partnership, so a pair fixed effect would absorb it; the teacher fixed effects do not, because many teachers appear in more than one partnership. Panel A's θ blends any return to duration with selection, since better teachers stay paired longer, while Panel B identifies θ from teachers observed in multiple partnerships of different completed lengths, holding each teacher's quality fixed. A causal reading of Panel B requires that, conditional on its two members and the controls, a partnership's realized duration be unrelated to shocks to its students' outcomes. That is, there is no selection on *match-specific* quality, which is precisely the object the table tests for.

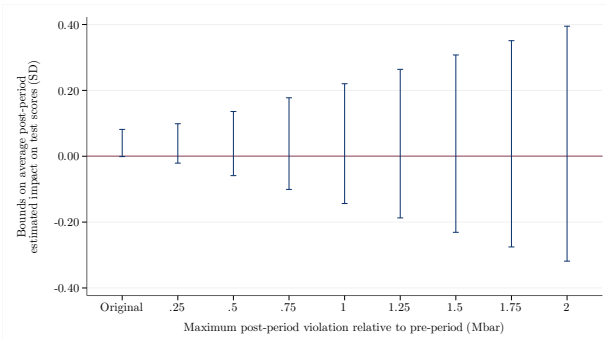
Figure A.20: Robustness to Violations of Parallel Trends



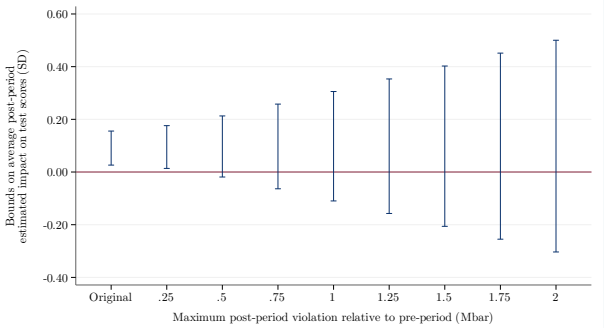
(a) Average effect, year 0



(b) Partner experience 16+, year 0



(c) Average effect, pooled post-periods



(d) Partner experience 16+, pooled post-periods

This figure presents Rambachan and Roth (2023) relative-magnitudes sensitivity analysis for the main event study (panels a, c) and the 16+ partner event study (panels b, d). The top row bounds the year-0 coefficient; the bottom row bounds the average of the three post-period coefficients, a more conservative target because admissible violations compound with horizon. Each interval is a 95% robust confidence set allowing post-period violations of parallel trends up to $\bar{M}$ times the largest pre-period violation; the leftmost interval in each panel is the conventional confidence interval. The average effect includes zero once any pre-period-scaled violation is admitted (both year 0 and pooled), as is mechanical under this exercise for a borderline-significant estimate; the 16+ partner effect excludes zero for $\bar{M} \leq 0.75$ (year 0) and $\bar{M} \leq 0.25$ (pooled). Computed with the conditional least-favorable hybrid method.

D.5 Selection into Co-Teaching Entry

An indicator for entering co-teaching the following year is regressed on the contemporaneous classroom score and its change,

$$\text{Enter}_{t,y+1} = \beta_1 \bar{\pi}_{ty} + \beta_2 \Delta\bar{\pi}_{ty} + \delta_y + \eta_{ty}, \tag{7}$$

estimated separately for 1–3-year and 4+-year teachers, with $\bar{\pi}_{ty}$ the teacher's standardized classroom mean and δ_y year effects; a companion logit of entry on classroom characteristics and year effects, estimated within each band, gives the AUC discrimination statistic reported

in Section 6.2. The joint F-test of the pre-period event-time coefficients, reported for the pooled main specification and separately by entry band, comes from rebuilding the stacked design and testing the pre-period ω_k jointly against zero. The remaining Section 6.2 contrasts are companion regressions on the same stacked sample: the entry-year difference in the probability of changing schools, the difference in assigned-partner experience for late versus early entrants, the late×pre-period interaction in the event study (the Ashenfelter dip), and the dip interacted with a 16+ experience partner within late entrants. The maintained requirement for the design is that entry be unrelated to the focal teacher's counterfactual *trajectory*. Insignificant β_1, β_2 in the 1–3 sample support this for early entrants; the dip, its concentration in 16+ partnerships, and the unstable late-band pre-trends are why 4+ entrants are excluded.

D.6 Student Fixed-Effects Specification (Figure A.16, Table A.6)

This replaces the prior-score controls in Equation (3) with a student fixed effect μ_i,

$$\pi_{ditsy} = \sum_{\substack{k=-3 \\ k\neq -1}}^{2} \omega_k \, YTC_{tyk} + \mu_i + \delta_{dt} + \gamma_{dy} + C'_{isy}\kappa + \varepsilon_{ditsy}, \tag{8}$$

where C are classroom-composition controls. In the stacked panel the student effect has no identifying variation, so each event is augmented with its students' observations from non-event classrooms in grades 3–8; these anchor rows carry their own event-time and event-teacher indicators and so pin down μ_i without entering the treated-control contrast. The unstacked panel version (Table A.6) runs the same comparison on all teachers' student-years. The identifying assumption is that, within student, the timing of a teacher's co-teaching entry is unrelated to that student's transitory score shocks; the post-entry years of co-teachers outside the analysis sample serve as a placebo.

D.7 Subgroup, Dose, and Functional-Form Variants of the Event Study

Several exhibits re-estimate Equation (3) on subsamples or with one added regressor, inheriting its identifying assumption: by partner experience, as bins and as polynomials of order one through four (Figures 6, A.10, 7); by the teacher's experience at entry (Figure 8); by partner degree (Figure A.11); adding an indicator for currently co-teaching to separate persistent from contemporaneous effects (Figure A.9); and comparing teachers who co-taught exactly n versus $n-1$ years to recover the marginal value of the nth year (Figure 9). The robustness panels that vary the prior-score lag structure and polynomial order (Figures A.12, A.13, A.14) and the sequential control build-up (Table A.8) change only the control set X_{psy}. For the partner-experience and entry-timing splits, the one requirement beyond the event study's assumptions is that variation in the splitting variable across treated teachers be unrelated to their counterfactual paths, which the bin-by-bin pre-trend checks in Section 6 address.

D.8 Imputation Estimator (Figure A.19)

The Borusyak et al. (2024) estimator fits the untreated-outcome model $\pi = \delta_t + \gamma_y + g_s + X'\beta$ on never- and not-yet-treated observations, imputes a counterfactual $\hat{\pi}^0$ for each treated observation, and averages $\pi - \hat{\pi}^0$ by event time. It is free of the non-convex weighting that can afflict stacked and two-way fixed-effects designs and accommodates the prior-score controls. Identification requires correct specification of the untreated-outcome model (parallel trends and no anticipation); the pre-period coefficients are placebo checks on it.

D.9 Event-Weighting Comparison (Table A.9)

The same stacked estimate is recomputed under three event weightings: the stacked design's implicit weights, equal weights across events, and weights proportional to each

event's student-year count. This is not a separate identification argument but a check that the headline estimate is not an artifact of how events are weighted; agreement across schemes is the object of interest.

D.10 Sensitivity to Parallel-Trends Violations (Figure A.20)

The event study identifies the effect off the assumption that treated and control teachers would have stayed on parallel paths. I relax that assumption using the Rambachan and Roth (2023) relative-magnitudes restriction, which allows post-period deviations from parallel trends up to $\bar{M}$ times the largest deviation observed in the pre-period and reports the breakdown value of $\bar{M}$ at which each effect loses significance. Figure A.20 traces the robust confidence sets, both for the year-zero coefficient and for the average of the three post-period coefficients.

The two effects behave very differently under this specific exercise, and the gap disciplines how much weight each can bear against trend violations. Within the relative-magnitudes restriction, the average effect has no margin: its conventional confidence interval already reaches zero, so admitting any pre-period-scaled violation renders it insignificant, both for the year-zero coefficient and for the pooled average. Any borderline $t \approx 2$ estimate would behave the same way, and pooling does not repair it. This is a statement about precision against parallel-trends violations, not about the estimate's stability: the average effect is essentially unchanged across weighting schemes, control sets, and the exclusion of school-movers, survives worst-case attrition trimming, and remains positive under the imputation estimator and the student fixed-effects strategy (significantly so in the better-powered unstacked comparison). The gain from highly experienced partners is sturdier on this dimension as well: it stays significant for violations up to $\bar{M} \leq 0.75$ times the largest pre-period violation for the year-zero coefficient, and the pooled version survives to $\bar{M} \leq 0.25$. I therefore treat the experienced-partner gradient as the finding that is robust on every dimension, including this most demanding one, and the policy analysis in Section 7 rests on it.

E Policy Simulation Details

This appendix provides additional detail on the assumptions underlying the policy simulation in Section 7.

E.1 Policy Structure

The simulation considers a policy where teachers with 2-3 years of solo teaching experience pair with highly experienced teachers (16+ years) for one year of co-teaching. After the co-teaching year ends, both teachers return to solo instruction. The early-career teacher carries any skill gains forward. The experienced teacher returns to their original role and remains available for future pairings.

The number of pairings each year is constrained by the supply of experienced mentors. Each mentor can support one pairing per year. I assume all available mentors participate and all eligible early-career teachers are paired if a mentor is available.

E.2 Parameter Values

Development effect. The early-career teacher gains $\delta = 0.10\sigma$ in effectiveness after co-teaching with a highly experienced partner. This estimate comes from Figure 6, which shows gains of approximately 0.10σ for teachers paired with partners who have 16+ years of experience.

Contemporaneous effect. Students in the co-taught classroom may experience direct benefits from having two teachers present. Jones and Winters (2024) estimate effects of approximately 0.02σ for students with disabilities and 0.01σ for students without. I use $\beta = 0.02\sigma$ in scenarios that include this benefit. Scenarios 3 and 4 set this to zero.

Opportunity cost. When the experienced teacher enters the co-taught classroom, their former solo classroom must be staffed by a replacement. I assume the replacement is a novice teacher. I set the effectiveness gap between a novice and a highly experienced teacher to $\gamma = 0.14\sigma$, the cumulative novice-to-veteran gap implied by the mathematics estimates of

Papay and Kraft (2015), the upper range of their results; their math–reading average, which the simulation's returns-to-experience function uses, implies a smaller gap of approximately 0.10σ between novice and 16-year teachers. Assuming the larger gap is conservative, because γ is the per-classroom opportunity cost of removing an experienced teacher from solo instruction and a larger γ raises the policy's cost.

Teacher turnover. Teachers leave the profession with probability $\tau = 0.20$ per year. This rate is consistent with national estimates of teacher attrition. The turnover rate determines both the steady-state experience distribution and the size of the mentor pool. With 0.20 annual turnover and no retention benefit, approximately 2.9% of teachers have 16+ years of experience.

Retention benefit. None is credited. Earlier versions of the simulation gave treated teachers lower turnover after their second post-treatment year, which enlarged the mentor pool over time. Appendix B finds no retention difference once both arms are held to the same survival requirement, so the parameter is set to zero in every scenario and the mentor pool never expands.

Skill decay. In Scenario 3, the development effect decays at 10% per year. A teacher who gained 0.10σ in year 1 retains 0.090σ in year 2, 0.081σ in year 3, and so on. This reflects the possibility that skills acquired through collaboration may fade without continued practice or reinforcement. Scenarios 1-2 assume no decay.

Returns to experience. Teachers improve with experience even without co-teaching. I use estimates from Papay and Kraft (2015) Table 2, averaging across math and reading. Teachers gain approximately 0.064σ over their first five years (0.0128σ per year), 0.031σ over years 5-15 (0.0031σ per year), and minimal gains thereafter (0.0003σ per year). These experience effects apply to all teachers regardless of treatment status. The simulation uses this returns-to-experience function to calculate baseline teacher productivity but does not explicitly credit the policy for shifting the experience distribution. Including this channel would increase the estimated benefits.

E.3 Steady-State Calculation

I simulate the policy forward until the distribution of teacher experience and treatment status reaches steady state. In each year:

1. Teachers with 16+ years of experience are assigned as mentors.
2. The share of mentors determines how many pairings occur.
3. Early-career teachers (2+ years of experience) are paired with mentors.
4. Treated teachers carry their skill gains (δ) forward, potentially with decay.
5. Teachers turn over at rates that depend on treatment status and time since treatment.
6. The workforce composition evolves based on these flows.

The simulation runs for 150 years to ensure convergence to steady state. Results report steady-state values after this convergence.

E.4 Benefit and Cost Calculations

Annual VA benefit. The simulation tracks the share of classrooms staffed by treated teachers and multiplies by the treatment effect (δ), adjusted for any skill decay. In scenarios with contemporaneous effects, it also adds the benefit from having two teachers present in co-taught classrooms (β times the mentor pool share).

Opportunity cost. Each pairing removes one experienced teacher from a solo classroom. The opportunity cost equals the mentor pool share (fraction of classrooms with mentors) multiplied by $\gamma = 0.14\sigma$.

Net annual effect. This equals total annual benefits minus annual opportunity cost, measured in units of student achievement (σ).

Lifetime earnings translation. I convert achievement effects to earnings using estimates from Chetty et al. (2014b). Their estimates imply that a one standard deviation

improvement in teacher value-added in a single grade raises a student's undiscounted lifetime earnings by approximately $39,000 (in 2010 dollars). Converting test-score effects to teacher value-added units using their middle school standard deviation of teacher effects (approximately 0.12σ of student test scores) and inflating to 2025 dollars with the CPI, each 0.01σ of test scores corresponds to roughly $5,000 of undiscounted lifetime earnings; equivalently, the simulation applies $500,000 per 1σ of net annual effect. The lifetime earnings per student equal the net annual effect times this conversion factor.

Years to positive cumulative. The simulation tracks cumulative net effects year by year. The "years to positive cumulative" is the first year where the sum of all past net effects becomes positive.

Break-even opportunity cost (γ^*). This is the opportunity cost that would make net annual benefits equal to zero: $\gamma^* = (\text{total benefits})/(\text{mentor pool share})$. Comparing this to the assumed $\gamma = 0.14\sigma$ shows how much margin for error the policy has.