Working Paper 2020–5
FALL 2020

Professional Development at Scale:

The Causal Effect of Obtaining an SEI Endorsement Under Massachusetts's RETELL Initiative

Jesse Bruhn, Nathan Jones, Yasuko Kanno, Marcus A. Winters

BU
Boston University Wheelock College of Education & Human Development
Wheelock Educational Policy Center

WEPC
WHEELOCK EDUCATIONAL POLICY CENTER

Professional Development at Scale: The Causal Effect of Obtaining an SEI Endorsement Under Massachusetts's RETELL Initiative

Abstract

We apply a difference-in-difference design to measure the effect of a teacher obtaining an endorsement in Sheltered English Immersion under Massachusetts's Rethinking Equity in the Teaching of English Language Learners initiative on students' learning outcomes. More than 35,000 in-service public school teachers completed the semester-long course. We find no effect on English learners' (ELs) average test scores, but modest positive spillovers for students with disabilities and other non-EL students. There are some improvements for ELs with very low English proficiency and students with very high or low prior math and reading scores. Training benefited teachers recently hired by their district.

Keywords: Professional Development, Teacher Quality, English Language Learners

1 Introduction

Professional development (PD) is perhaps the most common and longstanding strategy for improving the effectiveness of in-service teachers. American school systems annually spend about $18 billion on PD for in-service teachers (Gates 2014) despite little evidence that such training improves classroom effectiveness and no convincing evidence that it can do so at a meaningful scale. In a recent review and meta-analysis of research conducted in developed nations, Fryer (2017) reports that general PD focused on abstract concepts such as "rigor" or "classroom management" has a significant impact on teacher knowledge and practices but no impact on student outcomes, while more prescriptive managed PD that links training to specific curriculum and materials has significant and large impacts on student achievement. Yet, the evidence base is scant. Fryer's (2017) review included only five studies that employed a rigorous research design, several of which relied on small samples. Only five of the previously published studies evaluating the effect of PD for mathematics instruction met the What Works Clearinghouse evidence standards, and most of these studies report null findings (e.g., Borman et al. 2009; Garet et al. 2011, 2010; Santagata et al. 2010). The few existing studies that have used administrative data from large U.S. school systems have found generally null effects from exposure to PD for in-service teachers, but these analyses are limited to measuring the average effect of various PD as implemented by individual schools and districts across a large system without distinguishing between type of training, rather than evaluating the effectiveness of a particular PD to which a substantial number of teachers within a large system were exposed (Harris and Sass 2011; Jacob and Lefgren 2004).

We provide the first causal estimates for the impact of a statewide PD program in the United States. Specifically, we measure the effect of a teacher obtaining an endorsement in Sheltered English Immersion (SEI) under Massachusetts's Rethinking Equity in the Teaching of English Language Learners (RETELL) initiative on the performance of English learners (ELs) and other students they instruct. RETELL provides a unique example of a rigorous and time-intensive PD delivered at a nearly unprecedented scale. Over a 5-year period more

than 35,000 in-service public school teachers throughout Massachusetts completed a training equivalent in scope to a college-level semester-long course.

We apply a generalized difference-in-difference approach to estimate the causal effect of teachers' SEI endorsements on student performance. We find no overall effect of the training on ELs' average math and English language arts (ELA) scores. However, this null average effect appears to mask some heterogeneity associated with student and teacher attributes. We find some evidence that the training had a positive effect on teachers' impact for educators hired by the district within the previous three years —a proxy for classroom experience —but not for longer-serving teachers within the district. We find that the training had a positive impact on the standardized test scores of ELs with very low proficiency in English, but not for students nearer to reaching English proficiency. We also see some benefits from the training on ELs' progress on the test that annually assesses ELs' English language proficiency, though this result is restricted to ELs with relatively high English proficiency in the prior year. Further, we find some benefits from having a trained teacher for students who had very high or very low standardized math and reading scores in the prior year. Finally, we find a modest positive spillover on the average performance of students with disabilities and the larger group of non-EL students that the teacher instructs.

Our findings have direct policy implications within and beyond Massachusetts, with 24 states currently requiring or recommending EL-specific training for general education teachers (Education Commission of the States, 2019). Massachusetts provides an especially salient policy setting in which to estimate the effects of this rapidly growing requirement. Prior to the initiative, the vast majority of Massachusetts public school teachers with ELs in their classroom lacked sufficient training to instruct such students according to the state's educational model (Massachusetts Department of Elementary and Secondary Education [DESE], 2010). State policymakers initiated the training requirement in response to pressure from the U.S. Department of Justice, which argued that so few of the state's teachers had received adequate training to instruct ELs that it was in violation of federal law. The state developed the training in partnership with the federal government, thus ensuring that its content was

aligned with the federal standards and what the research considers to be best practices. Under RETELL, within a 5-year period the state moved from a situation in which relatively few general education teachers had EL-specific training to a policy that required any core academic teacher who instructs even a single EL to have acquired an SEI endorsement.

Our results contain both disheartening and encouraging elements for supporters of large-scale PD generally and specifically for training teachers to instruct ELs. On one hand, that the rapid expansion of teachers in the state who had received training in SEI instructional strategies did not lead to improvements in the average test scores of ELs is disappointing. If our results are considered entirely as a policy evaluation of the immediate impact of widespread SEI training under the RETELL initiative, we would have to conclude that the training requirement did not have the desired impact. On the other hand, our finding of a potential for positive impacts for recently hired teachers suggests that the training and SEI endorsement requirement could yeild benefits in the longer term as new teachers continue to enter the state's schools having completed the training during pre-service education. Further, the existence of spillovers for other students and the heterogeneous effects for teachers completing the training on the outcomes of both ELs and non-ELs suggests that the training did elicit improvements in teacher effectiveness, which is promising for the potential use of widespread PD as an intervention to improve teacher effectiveness. One plausible interpretation of our results is that widespread PD can improve teacher quality but that such policies might require specific targeting based on the success of the training for teachers and students with various characteristics.

Research demonstrating that teacher quality varies substantially and is an important determinant of student outcomes has sparked a variety of reforms and interventions within U.S. public schools (Rockoff 2004; Hanushek and Rivkin 2010; Chetty et al. 2014). Policies that focus on retaining more effective teachers have shown some promise (Dee and Wyckoff 2015; Adnot et al. 2017). But improving teacher quality at scale in a timely manner likely requires affecting current classroom teachers. Our findings for the impact of PD under the RETELL initiative add to the somewhat mixed results for the impact of efforts to improve

in-service teachers' effectiveness. For instance, recent evidence suggests that offering teachers higher salaries relative to alternative jobs increases the quality of entering teachers (Nagler et al. 2019), but uniform salary increases do not appear to increase the performance of teachers already in the classroom (de Ree et al. 2017) and linking teacher compensation to performance in U.S. public schools has yielded null or small effects (Springer et al. 2012; Goldhaber and Walch 2012; Glazerman and Seifullah 2010; Goodman and Turner 2011; Fryer Jr et al. 2012). However, in-service teachers appear to benefit from formal evaluations (Taylor and Tyler 2012) and coaching (Kraft et al. 2018).

In addition to improving our understanding of widespread PD, our study also contributes to the literature on teacher certification and endorsement. Though the initiative to endorse the state's current eligible teachers has reached its completion, the requirement that core academic teachers of ELs acquire an SEI endorsement remains for future teachers and will likely be fulfilled during pre-service training. Recent studies evaluating the efficacy of certification produced mixed results, though it appears that math certification likely has positive impacts (Harris and Sass 2011; Clotfelter et al. 2006, 2007, 2010; Goldhaber 2007; Sass 2015). The few recent quantitative studies finding an association between EL-specific training for general education teachers and ELs' higher achievement are correlational in nature and may not translate to the context of a statewide requirement (Loeb et al. 2014; Master et al. 2016; Betts and Zau 2003).[1]

The remainder of the paper proceeds as follows. Section 2 provides a basic description of EL education, the RETELL initiative, and SEI training. Section 3 describes the data. Section 4 describes our empirical strategy for estimating the causal effect of completing the training. We report the results in Section 5. In Section 6 we present a test of the plausibility of our identifying assumption for the analysis. Finally, Section 7 provides a brief summary of results and concludes.

2 Setting and Policy

2.1 Educating English Learners

ELs are among the most rapidly growing and lowest-performing student sub-populations in American public schools (e.g., National Education Association 2015; U.S. Department of Education 2018). ELs' relatively low academic outcomes are especially concerning because federal law requires schools to provide sufficient language support services to ensure that their access to content instruction is equivalent to that of native English speaking students. Under Title VI of the Civil Rights Act of 1964 and the Equal Educational Opportunities Act (EEOA) of 1974, states and school districts must "take 'affirmative steps' to address language barriers so that EL students may participate meaningfully in schools' educational programs" (U.S. Departments of Justice and Education, 2015). In *Lau v. Nichols* (1974), the seminal court case that definitively established the legal obligations of school districts and public schools to address ELs' language barriers, the U.S. Supreme Court declared: "There is no equality of treatment merely by providing students with the same facilities, textbooks, teachers, and curriculum; for students who do not understand English are effectively foreclosed from any meaningful education." The subsequent Castañeda v. Pickard (1981) ruling further specified that in order for a school district to be considered fulfilling its legal obligations, its EL program must meet three conditions: (a) that it is based on a sound educational theory; (b) that it is implemented with adequate resources and personnel; and (c) that over time it proves effective in eliminating ELs' language barriers.

Some of the legal obligations can be fulfilled by the language support services provided by English as a second language (ESL) and bilingual education specialist teachers. ESL teachers, for instance, might pull out ELs for English language development instruction or push into their general education classrooms to support ELs in academic instruction. However, concerns over segregating ELs from their English-speaking peers for separate instruction has led many states and school districts to favor an instructional model wherein ELs are kept in their regular classrooms as much as possible and general education teachers serve as both

content and language teachers (Harklau and Yang 2019). In such a model, general education teachers must be equipped to provide linguistic scaffolding along with academic instruction if ELs are to have equal access to academic content as non-ELs.

States have adopted policies requiring or recommending EL-specific PD for general education teachers in response to growing concern that ELs do not have sufficient access to instruction to meet their needs. Most ELs spend the majority of their school time in a general education classroom with a core academic teacher (Staehr Fenner 2013), and by some estimates more than half of public school teachers teach at least one EL (National Center for Education Statistics [NCES], 2012).[2] Much of the policy discussion regarding the lack of quality instruction for ELs has focused on the fact that ELs tend to be taught by inexperienced and less qualified teachers (Ballantyne et al. 2008; Cosentino de Cohen et al. 2005; Dabach 2015). But even otherwise fully qualified general education teachers may not have sufficient training to seamlessly incorporate the teaching of vocabulary and other language functions into their instruction to make academic content accessible to ELs and to foster their language development (Penner-Williams et al. 2017). A 2014 survey found that only 24% of elementary teacher education programs provided any training in EL-specific instructional strategies (Greenberg et al. 2015). Similarly, only 29.5% of general education teachers in U.S. public schools who have at least one EL in their classroom have had the opportunity to receive PD in EL education (NCES, 2012).

The logic underlying RETELL's SEI training requirement and similar, though less ambitious, initiatives in other states is to improve educational outcomes for ELs by ensuring that they are instructed by teachers sufficiently trained to meet their specialized needs. However, there is currently little to no empirical support for the effectiveness of such policies or for the training they require. Some recent studies find a positive association between EL-specific training and ELs' test scores (Loeb et al. 2014; Master et al. 2016; Betts and Zau 2003). However, none of these prior studies specific to EL certification and pre-service coursework employs a research design capable of leading to a causal estimate.

Notably, it is possible that the specific deficiencies in instruction for ELs targeted by

the SEI training and similar policies are not as daunting as commonly believed. Some recent studies have found that ELs' lower test scores relative to non-ELs are at least in part a reflection of other achievement gaps based on race/ethnicity and household income, thus challenging the conventional understanding that the so-called EL gap is driven by insufficient language services (Authors 2019; Callahan and Humphries 2016; Umansky et al. 2016). Indeed, using administrative data from Massachusetts during the time period of this study, [Authors] (2019) showed that controlling for student demographics and prior year test score nearly eliminates and in some cases reverses the difference in the test scores and educational attainment rates of ELs and non-ELs. These recent findings are relevant to the current work because if the relatively low performance of ELs is not primarily due to factors associated with their language proficiency or the quality of instruction that they receive relative to their classmates, then the potential for an intervention that targets improving practices for instructing ELs specifically might be limited.

2.2 Brief Recent History of EL Education in Massachusetts

In 2002, Massachusetts voters approved a ballot initiative requiring public school children to be taught in English language classrooms, which effectively eliminated transitional bilingual education programs in the state. Prior to this ballot initiative, 23% of ELs were enrolled in bilingual education. After Question 2 became law, the majority of these students were moved into SEI programs.

In theory, under the SEI model ELs were supposed to receive content instruction, mostly in English, from general education teachers who were trained to scaffold ELs' academic and language learning, while at the same time receiving more explicit English language development instruction from ESL specialist teachers. In practice, many ELs were instructed by teachers who were not trained in SEI practices. Although DESE mandated specialist training for ESL teachers, it did not mandate SEI training for general education teachers who had ELs in their classrooms. Instead, the state offered four categories of SEI training and encouraged general education teachers to undergo the training. Because the category

training was voluntary, only a small fraction of general education teachers took it. By 2010, 50,000 teachers, or 71% of the state's public school teachers, lacked training to work with ELs under the state's SEI model (DESE, 2010).

The statewide shortage of SEI-trained general education teachers meant that a large proportion of ELs were being taught by untrained teachers. For example, in 2010, only half of secondary-level ELs and a quarter of elementary-level ELs in Boston Public Schools, the largest school district in the state, received instruction from teachers who had either received category training or obtained an ESL license (U.S. Department of Justice [DOJ], 2011). Consequently, in July 2011, DOJ sent a letter informing the state that it was found in violation of the EEOA for its failure to require adequate training for SEI teachers. Referring to the state's decision not to mandate SEI training for general education teachers back in 2004, DOJ (2011) argued, "MADESE can no longer claim to be implementing an SEL Program model consistent with EEOA requirements if the voluntary PD program has resulted in a significant shortage of SEI teachers trained to educate ELL children in content classes seven years later" (p. 10).

2.2.1 RETELL

DESE responded to DOJ's concerns by proposing the RETELL initiative. At the heart of the initiative was a requirement that all core academic pre-service teachers and all core academic in-service teachers who were instructing ELs obtain an SEI teacher endorsement by June 30, 2016 (DESE, n.d.). Similarly, administrators (e.g., principals and supervisors) who evaluated SEI teachers were required to obtain the SEI administrator endorsement. SEI teachers could obtain the endorsement in one of three ways: (a) if they already possessed an ESL license, (b) passed an SEI test (SEI MTEL), or (c) completed a comprehensive SEI teacher endorsement course. Those who had no previous training or only the Category 1 training were required to take a full 45-hour endorsement course. Teachers who had completed three of the previously available category training were eligible to take a 15-hour short-bridge course, and those who had completed two of the category trainings could take a 24-hour long-bridge course (DESE,

2017). The policy applied to both new and existing teachers, and thus resulted in thousands of hours of PD across the state.

In this paper, we focus on measuring the effect of obtaining the SEI endorsement under RETELL on a teacher's contribution to student test score outcomes. However, it is worth noting that RETELL contained other aspects in addition to the training requirement. First, RETELL also required training for administrators where they learned how to provide supervision and support for teachers using SEI strategies. When the initiative began in 2013, the state also changed to the WIDA English development standards and began to use WIDA's ACCESS test to assess ELs' English language proficiency annually. By 2015 the RETELL initiative also included coaching and extended learning opportunities meant to build upon the SEI endorsement training. Thus, our analysis looks at only one important aspect of the full RETELL initiative.

After a small pilot meant to evaluate the content of the training (August and Haynes 2012), the statewide roll out of the requirement began in 2013. Due to the large number of teachers who were required to complete this training, the state assigned each district into one of three cohorts (DESE, 2017). The districts with the highest EL incidence and lowest EL academic performance were assigned Cohort 1. In other words, the districts with the highest needs for training received priority, whereas districts with fewer ELs and higher performing ELs received the training later (Chester 2012). The state funded the required PD during the 2- to 3-year cohort period. A teacher could choose when to begin the training but was required to complete the training within 1 year of beginning the course.

Only core academic teachers who were assigned to instruct ELs during the district's cohort years were required to obtain an SEI endorsement by the end of the cohort period and thus were eligible for the training. Unlike other PD opportunities, failure to successfully complete the training under RETELL resulted in meaningful consequences for teachers: Eligible teachers within a district who failed to acquire the endorsement by the end of the cohort period were unable to advance, renew, or extend their license, and they were required to earn the endorsement at their own expense.

2.3 SEI Professional Development Course

The overall purpose of the SEI endorsement is to help general education teachers develop proficiency in instructional strategies for making academic content accessible to ELs as well as scaffolding their English language development in the context of general education classrooms. The 45-hour SEI teacher endorsement course consists of 12 face-to-face sessions, 9 hours of online coursework, and a 2-hour small group capstone presentation. The course content is largely based on Margarita Calderón's (2007, 2011) Expediting Comprehension for English Language Learners (ExC-ELL) model, which emphasizes intentional lesson planning with goal setting, background building, and review of previously learned concepts; explicit instruction of reading, vocabulary, and writing; and promotion of ELs' active participation in classroom interactions. DESE lists 10 specific instructional strategies that must be taught and a set of readings that must be included in the course. DOJ's letter to DESE in 2011 included severe criticism of the inadequacy of the previous category training, especially in terms of the absence of explicit instruction on reading, writing, and academic vocabulary, as well as lack of opportunities for teachers to practice what they learned in a real classroom. The required SEI course thus has an explicit emphasis on those components. The course also involves at least four opportunities for participants to practice using some of the newly learned strategies in a real classroom and to reflect on their experiences with the course instructor and/or peers (DESE, 2016, 2017).

DESE hired and trained each of the instructors and evaluators. The course was administered by several approved venders across the state, including school districts and educational collaboratives in addition to colleges and universities. Teachers were required to pass the course, though they were not required to pass a particular test. About 3.8% of teachers who enrolled in the training failed at least once.

While some approved vendors of the SEI course voluntarily offered separate courses for elementary- and secondary-level teachers, there was no mandate from DESE to do so. Thus, most teachers received SEI training in a course that was supposed to apply for the entire K-12 education spectrum. Likewise, the course content included introduction to WIDA[3]

standards and assessments, therefore increasing teachers' awareness of what ELs should be able to do at each English language proficiency level. However, to what extent differentiated instruction for different language proficiency levels was incorporated into the training was left up to the individual instructors.

3 Data

We use longitudinal administrative data for the universe of Massachusetts public school students and their teachers for school years 2010-2011 through 2017-2018 provided by DESE. Student-level data include demographic and classification information and math and English language arts (ELA) scores on the state's spring standardized test – the Massachusetts Comprehensive Assessment System (MCAS) – which we standardize by subject, grade, and year to have a mean 0 and standard deviation 1. For ELs the data also include the student scores on the annual English proficiency test that is used to inform the decision to reclassify a student as no longer an EL.[4] We use data on student and teacher classroom assignments in order to match students to their teachers. See the Appendix for a detailed description of the data and matching process.

The estimation sample includes observations for students in Grades 4-8 and Grade 10 with valid information on included variables.[5] We exclude third grade students because the analysis controls for the student's test score in the prior year and testing begins in the third grade. The state does not administer a ninth grade test. For tenth grade students, we control for the student's test score in the eighth grade.

We include students that we can match to only a single teacher in the respective subject. Results are similar if we include students assigned to multiple teachers in a subject and we randomly choose the teacher to account for in the regression. We include only instructors who are classified as a "teacher," and thus we omit other classifications such as co-teachers and paraprofessionals. See the Appendix for a detailed description of the data construction and matching process.

We combine the administrative data with records from the SEI course, also provided by DESE. For each year from 2012-2013 (pilot year) through 2016-2017, the data identifies each teacher who enrolled in the SEI course, their completion date, course type (full, long-bridge, short-bridge, administrator), and indicates whether the teacher passed or failed the course.[6]

Table 1 presents descriptive statistics for relevant variables.

4 Identifying the Causal Effect of SEI Endorsement Training

The goal of this paper is to uncover the causal effect that obtaining an SEI endorsement under the RETELL initiative had on a teacher's contribution to student outcomes. Our ideal experiment would be to randomly assign the training to teachers after students have already been assigned to classrooms. However, as we discussed in section 2.2.1, the SEI endorsement training was rolled out to schools and districts on the basis of observables. Further, much of the training was conducted during school breaks and hence occurred prior to the mapping of students to classrooms.

We expect that a naive comparison of outcomes between students with trained and untrained teachers will be biased by two sources of selection. First, the roll out of the program across the state effectively prioritized teachers in urban and underperforming school districts. This source of selection would lead us to conflate student socioeconomic characteristics with the impact of completing the training. Thus we expect that the naive contrast is biased downward. Second, we worry that after teachers have received training, administrators within schools may endogenously sort students to trained teachers on the basis of ability. If trained teachers are more likely to receive EL or other typically low-performing students, this second source of selection would also lead the naive contrast to be biased downward.

We address these sources of selection empirically by leveraging cross-teacher variation in the timing of training via a generalized difference-in-difference strategy. Intuitively, we would like to compare the classroom-level trend in test score gains among teachers who have received

the training to those that have not. Because the comparison across teachers is via trends, our identification strategy addresses the first source of selection by effectively differencing out variation in the socioeconomic and demographic composition of students across schools and districts. To address the second source of selection, we will hold constant test scores from prior years and hence focus our attention on trends in test score *gains*. Thus we ask, "Do average test score gains increase suddenly in the classrooms of teachers that receive the training relative to the classrooms of teachers that do not?" Thus compositional changes emerging from the endogenous sorting of students to teachers are effectively controlled for by differencing out average student ability at the classroom level.

This intuition leads us to estimate variations of the following two-way fixed effects regression:

$$y_{it} = \alpha + \delta_j + \phi_t + \beta\tau_{jt} + \gamma y_{it-1} + \lambda X_{it} + \epsilon_{it} \tag{1}$$

Where y_{it} is the test score of student i at time t; δ_j is a fixed effect for teacher j, with $j = j(i,t)$ representing a one-to-one mapping between students and the teachers to whom they have been assigned at time t; ϕ_t is a time period fixed effect; τ_{jt} is an indicator for whether teacher j has completed the SEI endorsement training as of time t; X_{it} are controls for the demographic characteristics for student i at time t[7]; and ϵ_{it} is a projection residual which is uncorrelated with the included regressors by definition (Angrist and Pischke 2008).

Our parameter of interest is β. Without further assumptions, β identifies a weighted average of the underlying two-by-two classroom-level difference-in-difference comparisons (Goodman-Bacon 2018).[8] Provided the average classroom-level trend in test score gains of teachers who have not received or not-yet received the training is an accurate counterfactual for the average classroom-level trend of teachers who did receive training, β is properly interpreted as the causal effect that obtaining an SEI endorsement through the RETELL initiative has on student outcomes. This restriction on the treatment group counterfactual trend needed for causality is commonly referred to as a parallel trend assumption.

We acknowledge here that there is some subtlety to the interpretation of the estimand

of the two-way fixed effects model when there is variation in treatment timing. Rather than identifying the ATE, model 1 identifies a variance weighted average of any underlying heterogeneous treatment effects (Goodman-Bacon 2018). Of particular concern is the fact that the weights can be negative when the treatment effect changes over time (de Chaisemartin and d'Haultfoeuille 2020; Goodman-Bacon 2018). However, this is unlikely to be a problem in our application, since we find little evidence of strong dynamic effects in our event study type specifications. Further, we note that despite being a potentially biased estimate of the ATE, the variance weighted average can also be a substantially lower variance estimator (Angrist and Pischke 2008). Ultimately, the goal of this paper is to produce an estimate of the RETELL treatment effect that is as close as possible to the actual impact of the policy, in which case trading-in some bias in exchange for a reduction in variance can be desirable (Friedman et al. 2001).

Note that our preferred specification analyzes the data at the micro level rather than the (arguably more natural) specification which aggregates the data to the teacher level. This choice affords us two benefits: First, it transforms our estimate into a student-weighted average which we believe is the natural way to interpret the program impact; second, this allows us to leverage potential efficiency gains afforded by the inclusion of microlevel covariates. However, since training is assigned to teachers (not students), we cluster our standard errors at the teacher level to account for within-teacher autocorrelation in training status in accordance with the design-based view of Abadie et al. (2017).

We also note here that our choice to control for lagged test scores (and hence focus on gains) does impose a nontrivial cost. While this choice allows us to address the endogenous sorting of students to teachers, it will also force us to interpret our estimate as representing a pure short-run effect. To the extent that having an SEI endorsed teacher last year impacts a student's test scores in the current year, this causal effect of the training will also be indirectly controlled for by the inclusion of the student's test score from last year in the regression.[9] We address this concern in the Appendix by showing that having a trained teacher in the prior year does not have an independent effect on this year's test scores, conditional on the

student's prior-year test score. Thus, we believe the benefits of addressing the potential for endogenous sorting by focusing on test score gains outweighs the associated cost.

From a policy perspective, we are primarily interested in the effect of having a trained teacher for students who are currently classified as an EL, since this is the group of students that RETELL was specifically designed to impact. When the sample is restricted to include only current ELs, the regression nonparametrically controls for the student's proficiency in English at the end of the previous year as measured by their performance level on the proficiency assessment that the state uses to inform reclassification decisions.[10] When the sample is restricted to include Ever ELs (i.e., students who we observe at any prior year to be classified as EL), we control for the student's prior proficiency level if it is available. These controls account for differences in the student's ability to understand English, which is part of the data-generating process for student math and ELA test scores. Results in the Appendix show that the estimates are nearly unchanged when this variable is excluded.

4.1 Spillovers and Heterogeneity

Though RETELL was specifically intended to improve outcomes for ELs, there is considerable potential for spillover effects for non-ELs. We thus look for spillovers in treatment effect by estimating Equation (1) on samples restricted to include one of several student subgroups of interest. We measure the effect of having a trained teacher on the performance of language minority students generally by limiting the sample to include Ever-ELs. We also estimate models restricted to students who have a disability and thus receive special education services because some instructional practices recommended in the EL literature, such as modeling and explicit instruction (e.g., August 2007; Goldenberg 2008), align with practices recommended for struggling readers (Connor et al. 2011) and students with disabilities (Authors 2014). Further, given some recent evidence from Cohen (2018) that the explicit, systematic instruction recommended in the EL literature is associated with math and reading achievement for the general student population, we also measure the effect of the training on the performance of the larger group of students who we never observe being

classified as an EL (Never ELs).[11]

In addition to estimating average effects for ELs and other subgroups, we present models looking for heterogeneity in the treatment effect by a variety of student and teacher characteristics. We measure whether the effect differs for teachers who completed the full training and those who completed one of the shorter supplemental versions of the training.[12] Due to the large demographic differences across cohorts and the potential that the training could change over time, we look for differential effects by cohort and by the year that the teacher completed the training. We also measure whether the effect differs by grade level or according to the student's score on the math or ELA exam at the end of the prior year. For ELs, we evaluate whether the effect differs for students with different levels of proficiency in English. Finally, we consider whether the effect of completing the training differs according to the number of years since the teacher was hired into the school district, which we treat as an imperfect proxy for years of teaching experience. When estimating models looking for heterogeneous impacts by teacher and student attributes, we run single models that interact treatment with the characteristic value. When such analyses lead us to make multiple comparisons, we use the Bonferroni correction when inferring differences between groups.

5 Results

5.1 Average Effect of Training on Student Math and Reading Scores

Table 2 reports the results from regressions evaluating the effect of completing the SEI endorsement training on teachers' impact on average student ELA and math scores. For each subgroup, the table reports results from models that evaluate the effect of completing the training overall as well as a model that differentiates according to whether the teacher received the full training or one of the shorter versions. Though at first glance there appear to be some differences according to the type of training, the estimates by training type are

less precise and are not qualitatively different than the main results. Thus, for the remainder of the paper we report results only from models that combine all forms of training.[13]

For ELs, our subgroup of primary interest, we find no significant effect from completing the training overall in either ELA or math. For both subjects the estimated effect of the training is negative and the coefficients are estimated precisely enough to detect meaningful effects. The lone significant effect finds that completing the short-bridge version of the training decreased the teacher's impact on the math scores of ELs in their classroom. The results from models that include all students who are observed as an EL at some point in the data (Ever ELs) are quite similar to the results from models restricted to current ELs.

We find some evidence of positive spillover effects for students with disabilities and for the larger group of Never ELs on the ELA exam overall and on the math exam for those who completed the full training. Though statistically significant, the magnitude of these potential spillover effects on the ELA exam is modest.

5.2 Heterogeneous Effects on Student Math and Reading Scores

5.2.1 Effect by Cohort and Year Training Completed

We begin our analysis of differential effects by considering whether the impact of being assigned to a trained teacher varies across cohorts or the year that the teacher completed the training. We can separately evaluate these effects because cohort periods overlapped.

Figure 1 illustrates the results from models that separately evaluate the effect of completing the training by cohort. There appear to be some across-cohort differences in the effect, but few are statistically significant even before adjusting for multiple comparisons and none are large enough to meaningfully change the interpretation of the main effect described in the prior section. For current ELs, we find no significant impact from completing the training for any of the cohorts in either subject. It appears that the math performance of Ever ELs in the second and third cohorts may have benefited from their teachers completing the training. Finally, the positive average spillover effect for students with disabilities and

Never ELs on the ELA exam shown on Table 2 appears to be driven by teachers in the first cohort.

The slight potential differences in the effectiveness of the training across cohorts might be due to some combination of differences in the characteristics of students in the cohorts and differences in the timing in which the training was completed. Figure 2 considers the latter case by illustrating the results from models that look for a differential impact according to the year that the teacher completed the training.

Similar to the across-cohort results, differences in the effects across completion years are rarely statistically significant even before adjusting for multiple comparisons and are not large enough to justify a reinterpretation of the main result. Completing the training in 2013 (the first statewide year) had a significant negative effect on the ELA performance of ELs in a teacher's class, but this is not significantly different from the estimated effects for the following two years. The effects of the training for both Ever ELs and Never ELs appear quite stable over time. The positive effect of the training for students with disabilities on the ELA exam appears to wane beginning in 2016, though the potential differences are not significant after accounting for multiple comparisons.

5.2.2 Effect by Grade Level

Figure 3 illustrates the coefficient and 95% confidence interval for the impact of completing the training by grade level for each subgroup. For each subgroup we estimate the effect by grade in a single regression that is equivalent to Equation (1) but replaces τ_{jt} with separate treatment indicators for each grade level.

For ELs, we find no statistically significant effect in any grade on the ELA exam, and only in the sixth grade on the math exam. It seems likely that the negative effect on the sixth grade math exam is spurious given the large number of comparisons and that the estimate does not fit a clear pattern of results. The pattern of results on the math exam is generally similar across grades. On the ELA exam, training tends to have a more negative effect in earlier grades than in the later grades. However, the apparent differences by grade on the

ELA exam are not statistically significant even before we adjust for multiple comparisons.

The modest positive average spillover effects for students with disabilities on the ELA exam reported in Table 1 appear to be primarily driven by students in grades four through six. The effect for Never ELs appears to be quite similar across grades, with the exception of students in grade 4. However, in these cases the differences in the effect across grades is not statistically significant.

5.2.3 Effect by Prior MEPA/ACCESS Score

Figure 4 illustrates the results from models evaluating whether the effect of RETELL training on ELs differed according to the student's proficiency level in English. The sample for these regressions includes only students currently classified as an EL. The models are equivalent to Equation (1) but replace τ_{jt} with separate treatment indicators for the level students scored on the test assessing their English proficiency. Students who scored in Level 1 have little to no proficiency in English, while those scoring Level 5 are nearly English proficient. In addition to the overall effect, the figure illustrates results from models restricted to include students in Grades 4 and 5 and for students in Grades 6-8, and Grade 10, combined.

Students at or above Level 3 English proficiency did not appear to benefit from their teacher receiving the training. However, we find some evidence that the training benefits students with very little proficiency in English on the ELA exam. Level 1 students experienced a substantial estimated increase of 0.098 standard deviations on the ELA exam if their teacher had completed the training. The effect in ELA is most pronounced for students in higher grade levels, though the reduction in the observations from splitting the sample by both grade level and English proficiency level lead to quite imprecise estimates. After adjusting for multiple comparisons, the differences between Level 1 and Level 5 effects remain statistically significant at the 5% level of confidence, while the difference between Level 1 and the remaining three levels are significant at the marginal 10% level.

5.2.4 Effect by Teacher's Years of Employment within District

Training might impact more and less experienced teachers differently. Unfortunately, our data does not include specific information on the teacher's prior years of classroom experience. However, we do observe each teacher's first hire date within the district, which we argue serves as an imperfect but reasonable proxy for years of experience. We bin teachers according to years since first hired within the district into categories that largely represent quartiles for the workforce in Massachusetts.

The results illustrated on Figure 5 suggest that the training had a significant positive effect on the performance of teachers hired within the previous three years. For ELs, the estimate is positive but imprecisely estimated on the ELA exam. The effect for ELs is statistically significant on the math test for teachers in secondary grades. The magnitude of the effect of the training for these newly hired teachers when combined across grades (between 0.03 and 0.06 standard deviations) is quite consistent across subjects, grade level, and subgroup. In contrast, we find no evidence that the training had an impact on the performance of teachers who have been working within the district for longer than three years. The imprecision in the ELA analysis leads all comparisons across categories for years since hired to be statistically insignificant. However, several of the differences between newly hired teachers and teachers with more in-district experience remain statistically significant at the 10% level after applying the Bonferoni adjustment for multiple comparisons.

5.2.5 Effect by Prior Student ELA and Math Score

Figure 6 illustrates the results from models that evaluate whether the average effect of the training differed according to a student's math or ELA score at the end of the previous year. We separate students into bins according to the decile of the student's score in the subject in the prior year on the overall test score distribution. Thus, the test scores represented in each decile bin are the same regardless of whether the model is evaluating the sample of ELs or of the other non-EL categories. As a result, very few ELs are represented in the top decile categories, as the majority of ELs have below-mean test scores, which can be seen in

the expansion of the confidence intervals for higher grade levels.

The effect of the training is similar for students at most points on the prior test score distribution. However, there is a consistent statistically significant positive effect for students in both the bottom and top deciles of prior score within the subject. The difference between the effect for students in the bottom decile and each other decile remains statistically significant after adjusting for multiple comparisons. In some cases the magnitude of the effect is substantial.

5.3 Effect on ELs' Progress Toward English Language Proficiency

Table 3 reports the results from models evaluating the impact of training completion on a teacher's influence on an EL's progression toward proficiency in English according to the assessment used for reclassification decisions. These models are identical to Equation (1) but the dependent variable is an indicator for whether the student's EPL on the test increased by at least one level from the prior year. Also, the model does not control for the student's prior ELA or math score. The sample for these models includes students in all grades K-12 rather than only students in grades that administer standardized tests in math and ELA. To put the estimates into context, as indicated in the bottom of the table, about half of students increase their EPL by at least one level each year. The table reports results overall and separated by the student's proficiency level the prior year.

Overall, we find that completing the training does not significantly affect the probability that ELs in the teacher's classroom improve by at least one level on the English proficiency assessment on average. However, that average finding masks heterogeneity in the effect based on the student's proficiency level at the end of the prior year. Students with very low proficiency in English were less likely to improve by at least one level if their teacher completed the training. However, students who the prior year had high enough proficiency in English to score in Level 4 were substantially more likely to improve by at least one level if their teacher were trained.[14]

6 Identification Test

The primary assumption required to interpret β as the causal effect of completing the training on the outcomes of a teacher's students is that there are no time-variant factors that are associated with both the timing of the teacher completing the training and the outcomes of students in her class. We test the plausibility that this assumption holds by conducting an event study analysis that measures changes in teacher effects in the years leading up to and following the training. In particular, we estimate a regression taking the form:

$$y_{ijt} = \alpha + \gamma y_{ijt-1} + \lambda X_{ijt} + \delta_j + \phi_t + \sum_{k=-3}^{3} \beta_k D_{jk} + \epsilon_{ijt} \tag{2}$$

where k is an index for the number of years from the teacher's training, such that $k = 0$ represents the year prior to completing the training. We estimate Equation (2) for each subgroup and for a variety of different categorizations.

Figure 7 illustrates results from models looking at the impact of the training on the average test scores for ELs, students with disabilities, and students never observed to be ELs for the full sample and for samples restricted by grade level. Consistent with the results in Table 1, there is no clear post-training difference in the outcomes for ELs but there appear to be positive post-treatment effects for students with disabilities and students never observed as ELs.

The patterns illustrated on the figures are generally consistent with applying a causal interpretation to the significant impacts identified in the regressions. In the cases where we previously found impacts from the training – the ELA exam for students with disabilities and Never ELs – the pattern is for there to be no significant difference in years prior to completing the training followed by a jump in performance in the year immediately following the training.

7 Summary and Conclusion

We use a generalized difference-in-difference design to estimate the causal effect of in-service core academic public school teachers in Massachusetts completing an intensive PD in order to obtain an SEI endorsement under the state's RETELL initiative. Our study represents the first causal estimates for the impact of a rigorous and time-intensive PD for public school teachers in a developed nation conducted at a statewide scale.

The purpose of the SEI endorsement requirement under the RETELL initiative was to increase the educational outcomes for ELs across the state. We find no evidence that the training had a significant effect on a teacher's impact on the math or ELA scores of ELs in their classroom, on average. Thus, to the extent that the SEI endorsement requirement was meant to have a broad impact on ELs' academic performance across grades and subjects, our results suggest that it did not achieve the intended effect.

However, the null average effect masks potentially important heterogeneity in the treatment impact. Teachers recently hired by the district benefited from the training, though teachers who had been employed within the district for more than three years did not. Further, a teacher's training completion had a positive effect on ELs in their classroom who had very low proficiency in English, but not on students who were closer to English proficiency. That the positive effect of the training was limited to only a subset of teachers and ELs has important implications for our view of the SEI endorsement requirement under the RETELL initiative and for states considering similar policies in the future. On one hand, our results suggest that an intensive PD distributed to teachers at an expansive scale can lead to teacher quality improvements for some teachers. On the other hand, the specific nature of the effect suggests a strategy of first piloting the training on a representative group of teachers in order to better understand its effects and then targeting the training to those who are likely to benefit from it, rather than a costly universal mandate.

Our finding that the positive effect of the training was restricted to teachers recently hired within the district has implications for PD's potential as a tool for making widespread improvements in the quality of the current teacher labor force. Our result is consistent with

Lu et al. (2019), which similarly found a positive effect of PD that was restricted to only "less qualified" teachers in the context of a developing country. If PD effects are limited to recently hired teachers, then it stands little chance of improving teacher quality at a large scale. However, the finding raises the question of whether it would be more efficient to embed the techniques developed in the PD within pre-service teacher training rather than provided as PD during the teacher's first several years of employment. If the training has a positive effect on early-career teachers, then the gains in student outcomes may be delayed but substantial as new teachers enter the labor force and eventually become the more experienced teachers within the system.

When interpreting our results, it is important to keep in mind that our analysis measures the causal effect of completing the SEI endorsement training under the RETELL initiative, which is not necessarily the same as the effect of utilizing the strategies that are taught in the training. We do not observe the extent to which teachers changed their instructional practices after receiving the training. Further, it is feasible that the expansive nature of the training across the state affects the quality and consistency of the training. Future research using causal research designs is necessary to more fully understand the impact of the instructional strategies that were part of the training on a teacher's impact on the academic performance of ELs and other students.

Further, our analysis is only able to measure the effect of completing the training on the impact that teachers have on student test scores in the short-run. It is possible that teachers could continue to improve over time due to the training. In addition, we might expect that having trained teachers throughout schooling could manifest in long-run outcomes such as educational attainment.

Nonetheless, the presence of spillovers from the training for students with disabilities and for the larger group of students never observed as ELs has potentially important implications for the teacher training literature more generally. This result suggests that there are at least aspects of the training designed to improve instruction for ELs that are associated with more effective teaching generally.

Endnotes

1. In the only prior study that attempted to measure the effect of requiring universal EL training of which we are aware, López et al. (2013) examined the relationship between state requirements for teacher education across the U.S. and Hispanic ELs' fourth grade National Assessment of Educational Progress (NAEP) reading scores. They found that a state policy requiring all teachers to receive EL-specific training was associated with lower reading achievement, whereas a policy requiring only teachers who teach ELs to have an ESL/bilingual education certification was associated with higher gains. Though the study is notable in its attempt to examine the effect of policies for universal EL training on student achievement across the United States, its basic comparison of test score outcomes controlling for observed characteristics of students across states is far too limited to plausibly lead to a causal interpretation of the treatment effects.

2. See the NCES Schools and Staffing Survey (SASS) here:
`https://nces.ed.gov/surveys/sass/tables/sass1112_498_t1n.asp`

3. WIDA (https://wida.wisc.edu/) is a consortium that provides English language proficiency (ELP) standards, assessment tool, and PD in EL education. The WIDA ELP standards have been adopted by 40 states, including Massachusetts.

4. Prior to 2012 the test was the MEPA. The state converted to the ACCESS test for 2012 through 2016, and then converted to the ACCESS 2.0 in 2017. We use a tool available from by DESE to convert ACCESS and ACCESS 2.0 scores to the MEPA scale for comparability.

5. Results for the later grade analysis are similar if we remove 10th grade observations.

6. We also observe the course instructor, but we do not observe the institution that provided the training.

7. The student characteristics include gender, free/reduced lunch status, special educa-

tion classification, race (White, Black, Asian, Indian/Alaskan, and/or Hawaiian), and Hispanic ethnicity.

8. Strictly speaking, this statement is only true if we first aggregate the data to the teacher level and run the natural analogue to Equation (1). Since Equation (1) is at the student level, we will instead recover a student-weighted version of the average effect from the aggregate regression.

9. Some readers may be concerned that the choice to classify a student as EL may itself be an outcome of having a trained teacher, leading to a "bad control" problem that could cause our subsample estimates to exhibit bias. However, such bias is very unlikely because a student's initial EL classification is entirely determined by their performance on a language assessment test taken within 30 days of entering the school and is thus independent of the teacher.

10. There are five discrete MEPA levels. Students that score a level five, are then re-classified as non-EL.

11. We are aware that in the literature, Ever ELs have come to mean students who have been identified as ELs at some point in their K-12 schooling regardless of their later reclassification status and Never ELs to mean students (monolingual or multilingual) who have never been identified as ELs since entering school (e.g., Kieffer 2018; Umansky and Thompson 2017). However, in this study, we are using Ever ELs to refer to students whom we observe to have been identified as ELs in at least one grade in our dataset (i.e., fourth grade on) and Never ELs to refer to students who are not identified as ELs in our dataset.

12. We remove from the sample teachers who completed the administrator version of the training. These observations likely represent administrators who also teach one or more courses. Estimates for the effect of this sort of training are quite imprecise, and removing this small number of teachers has no meaningful effect on the findings.

13. As shown in the Appendix, the results from each analysis of treatment heterogeneity are also qualitatively similar when we remove all teachers who completed one of the versions other than the full training.

14. The pattern of results on the English proficiency assessment and the ELA exam are quite different, but are not necessarily contradictory. First, the exams are different. But more importantly, the ELA test analysis measures student test score gains, while the analysis of the English proficiency exam measures the probability of moving forward at least one performance level at the end of the year. Unfortunately, because the state changed the English proficiency assessment twice during our sample period, it is not possible to accurately measure student gains on the assessment, though the state gives guidance on equating performance levels across the assessments.

References

A. Abadie, S. Athey, G. W. Imbens, and J. Wooldridge. When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research, 2017. URL https://www.nber.org/papers/w24003.pdf.

M. Adnot, T. Dee, V. Katz, and J. Wyckoff. Teacher turnover, teacher quality, and student achievement in dcps. *Educational Evaluation and Policy Analysis*, 39(1):54–76, 2017. URL https://doi.org/10.3102/0162373716663646.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

D. August and C. Haynes, Erin & Paulsen. An evaluation of the massachusetts pilot teachers' sei endorsement course. 2012.

T. August, Diance & Shanahan. *Developing Reading and Writing in Second-Language Learners: Lessons from the Report of the National Literacy Panel on Language-Minority Children and Youth*. Abingdon, UK: Routledge, 2007.

Authors. 2014.

K. G. Ballantyne, A. R. Sanderman, and J. Levy. Educating english language learners: Building teacher capacity. 2008. URL https://files.eric.ed.gov/fulltext/ED521360.pdf.

J. R. Betts and L. Zau, Andrew & Rice. *Determinants of student achievement: New evidence from San Diego*. Public Policy Institute of California, 2003. URL https://www.ppic.org/content/pubs/report/R_803JBR.pdf.

K. M. Borman, B. A. Cotner, R. S. Lee, and R. Boydston, Theodore L & Lanehart. Improving elementary science instruction and student achievement: The impact of a professional development program. 2009.

R. M. Callahan and M. H. Humphries. Undermatched? school-based linguistic status, college going, and the immigrant advantage. *American Educational Research Journal*, 53(2):263–295, 2016. URL https://doi.org/10.3102/0002831215627857.

D. Chester. Memorandum to members of the board of elementary and secondary education, 2012.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9):2633–2679, 2014. URL https://doi.org/10.1257/aer.104.9.2633.

C. T. Clotfelter, H. F. Ladd, and J. L. Vigdor. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, 41(4):778–820, 2006. URL https://doi.org/10.3368/jhr.XLI.4.778.

C. T. Clotfelter, H. F. Ladd, and J. L. Vigdor. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6):673–682, 2007. URL https://doi.org/10.1016/j.econedurev.2007.10.002.

C. T. Clotfelter, H. F. Ladd, and J. L. Vigdor. Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects. *Journal of Human Resources*, 45(3):655–681, 2010. URL https://doi.org/10.3368/jhr.45.3.655.

J. Cohen. Practices that cross disciplines?: Revisiting explicit instruction in elementary mathematics and english language arts. *Teaching and Teacher Education: An International Journal of Research and Studies*, 69(1):324–335, 2018. URL https://doi/org/10.1016/j.tate.2017.10.021.

C. M. Connor, F. J. Morrison, C. Schatschneider, J. R. Toste, E. Lundblom, E. C. Crowe, and B. Fishman. Effective classroom instruction: Implications of child characteristics by reading instruction interactions on first graders' word reading achievement. *Journal of Research on Educational Effectiveness*, 4(3):173–207, 2011. URL https://doi.org/10.1080/19345747.2010.510179.

C. Cosentino de Cohen, N. Deterding, and B. C. Clewell. Who's left behind? immigrant children in high- and low-lep schools. 2005. URL https://www.fcd-us.org/assets/2016/04/WhosLeftBehind.pdf.

D. B. Dabach. Teacher placement into immigrant english learner classrooms: Limiting access in comprehensive high schools. *American Educational Research Journal*, 52(2):243–274, 2015. URL https://doi.org/10.3102/0002831215574725.

C. de Chaisemartin and X. d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96, 2020. URL https://doi.org/10.1257/aer.20181169.

J. de Ree, K. Muralidharan, M. Pradhan, and H. Rogers. Double for nothing? Experimental evidence on an unconditional teacher salary increase in Indonesia. *The Quarterly Journal of Economics*, 133(2):993–1039, 11 2017. ISSN 0033-5533. doi: 10.1093/qje/qjx040. URL https://doi.org/10.1093/qje/qjx040.

T. S. Dee and J. Wyckoff. Incentives, selection, and teacher performance: Evidence from impact. *Journal of Policy Analysis and Management*, 34(2):267–297, 2015. URL https://doi.org/10.1002/pam.21818.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics, 2001.

R. G. Fryer. The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments*, volume 2, pages 95–322. Elsevier, 2017. URL https://doi.org/10.1016/bs.hefe.2016.08.006.

R. G. Fryer Jr, S. D. Levitt, J. List, and S. Sadoff. Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. Technical report, National Bureau of Economic Research, 2012. URL https://www.nber.org/papers/w18237.pdf.

M. S. Garet, A. J. Wayne, F. Stancavage, J. Taylor, K. Walters, M. Song, S. Brown, S. Hurlburt, P. Zhu, S. Sepanik, and E. Doolittle, Fred & Warner. Middle school mathematics professional development impact study: Findings after the first year of implementation. 2010. URL https://ies.ed.gov/ncee/pubs/20104009/pdf/20104009.pdf.

M. S. Garet, A. J. Wayne, F. Stancavage, J. Taylor, M. Eaton, K. Walters, M. Song, S. Brown, S. Hurlburt, P. Zhu, and o. SEpanik. Middle school mathematics professional development impact study: Findings after the second year of implementation. 2011. URL https://ies.ed.gov/ncee/pubs/20114024/pdf/20114024.pdf.

M. Gates, B & Gates. Teachers know best: Teachers' views on professional development. *Bill and Melinda Gates Foundation*, 2014. URL https://k12education.gatesfoundation.org/download/?Num=2336&filename=Gates-PDMarketResearch-Dec5.pdf.

S. Glazerman and A. Seifullah. An evaluation of the teacher advancement program (tap) in chicago: Year two impact report. 2010.

C. Goldenberg. Teaching english language learners: What the research does–and does not–say. 2008.

D. Goldhaber and J. Walch. Strategic pay reform: A student outcomes-based evaluation of denver's procomp teacher pay initiative. *Economics of Education Review*, 31(6):1067–1083, 2012. URL 10.1016/j.econedurev.2012.06.007.

E. Goldhaber, Dan & Anthony. Can teacher quality be effectively assessed? national board certification as a signal of effective teaching. *The Review of Economics and Statistics*, 89 (1):134–150, 2007. URL https://doi.org/10.1162/rest.89.1.134.

S. Goodman and L. Turner. Does whole-school performance pay improve student learning? *Education Next*, 11(2):66–72, 2011.

A. Goodman-Bacon. Difference-in-differences with variation in treatment timing. Technical

report, National Bureau of Economic Research, 2018. URL https://www.nber.org/papers/w25018.pdf.

J. Greenberg, K. Walsh, and A. McKee. 2014 teacher prep review: A review of the nation's teacher preparation programs, 2015.

E. A. Hanushek and S. G. Rivkin. The quality and distribution of teachers under the no child left behind act. *Journal of Economic Perspectives*, 24(3):133–50, 2010.

L. Harklau and A. H. Yang. Educators' construction of mainstreaming policy for english learners: A decision-making theory perspective. *Language Policy*, 19:87–110, 2019.

D. N. Harris and T. R. Sass. Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8):798–812, 2011. URL https://doi.org/10.1016/j.jpubeco.2010.11.009.

B. A. Jacob and L. Lefgren. Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, 86(1):226–244, 2004. URL https://doi.org/10.1162/003465304323023778.

K. D. Kieffer, Michael J & Thompson. Hidden progress of multilingual students on naep. *Educational Researcher*, 47(6):391–398, 2018. URL https://doi.org/10.3102/0013189X18777740.

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

S. Loeb, J. Soland, and L. Fox. Is a good teacher a good teacher for all? comparing value-added of teachers with their english learners and non-english learners. *Educational Evaluation and Policy Analysis*, 36(4):457–475, 2014. URL https://doi.org/10.3102/0162373714527788.

F. López, M. Scanlan, and B. Gundrum. Preparing teachers of english language learners: Empirical evidence and policy implications. *Education Policy Analysis Archives*, 21:20, 2013.

M. Lu, P. Loyalka, Y. Shi, F. Chang, C. Liu, and S. Rozelle. The impact of teacher professional development programs on student achievement in rural china: evidence from shaanxi province. *Journal of Development Effectiveness*, 11(2):1–27, 2019. URL https://doi.org/10.1080/19439342.2019.1624594.

B. Master, S. Loeb, C. Whitney, and J. Wyckoff. Different skills? identifying differentially effective teachers of english language learners. *The Elementary School Journal*, 117(2): 261–284, 2016.

M. Nagler, M. Piopiunik, and M. West. Weak markets, strong teachers: Recession at career start and teacher effectiveness. *Journal of Labor Economics*, 38(2):453–500, 2019. URL https://doi.org/10.1086/705883.

National Education Association. Understanding the gaps: Who are we leaving behind–and how far? 2015. URL https://www.nea.org/assets/docs/18021-Closing_Achve_Gap_backgrndr_7-FINAL.pdf.

J. Penner-Williams, E. I. Díaz, and D. G. Worthen. Plcs: Key pd component in learning transfer for teachers of english learners. *Teaching and Teacher Education*, 65:215–229, 2017.

J. E. Rockoff. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2):247–252, 2004.

R. Santagata, N. Kersting, and J. W. Givvin, Karen B & Stigler. Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. *Journal of Research on Educational Effectiveness*, 4(1):1–24, 2010. URL https://doi.org/10.1080/19345747.2010.498562.

T. R. Sass. Licensure and worker quality: A comparison of alternative routes to teaching. *The Journal of Law and Economics*, 58(1):1–35, 2015.

M. G. Springer, D. Ballou, L. Hamilton, V. Le, J. Lockwood, D. McCaffrey, M. Pepper, and B. Stecher. Final report: Experimental evidence from the project on incentives in teaching (point). 2012.

D. Staehr Fenner. Implementing the common core state standards for english learners: The changing role of the esl teacher. [proceedings]. february 2013 tesol international association convening. *TESOL International Association*, 2013. URL `https://www.tesol.org/docs/default-source/advocacy/ccss_convening_final-8-15-13.pdf?sfvrsn=68590cdc_10`.

E. S. Taylor and J. H. Tyler. The effect of evaluation on teacher performance. *American Economic Review*, 102(7):3628–3651, 2012.

I. M. Umansky and G. Thompson, Karen D & Díaz. Using an ever–english learner framework to examine disproportionality in special education. *Exceptional Children*, 84(1):76–96, 2017.

I. M. Umansky, R. A. Valentino, and S. F. Reardon. The promise of two-language education. *Educational Leadership*, 73(5):10–17, 2016.

U.S. Department of Education. Educational experiences of english learners. 2018.

Figures

Figure (1) Effect by Cohort

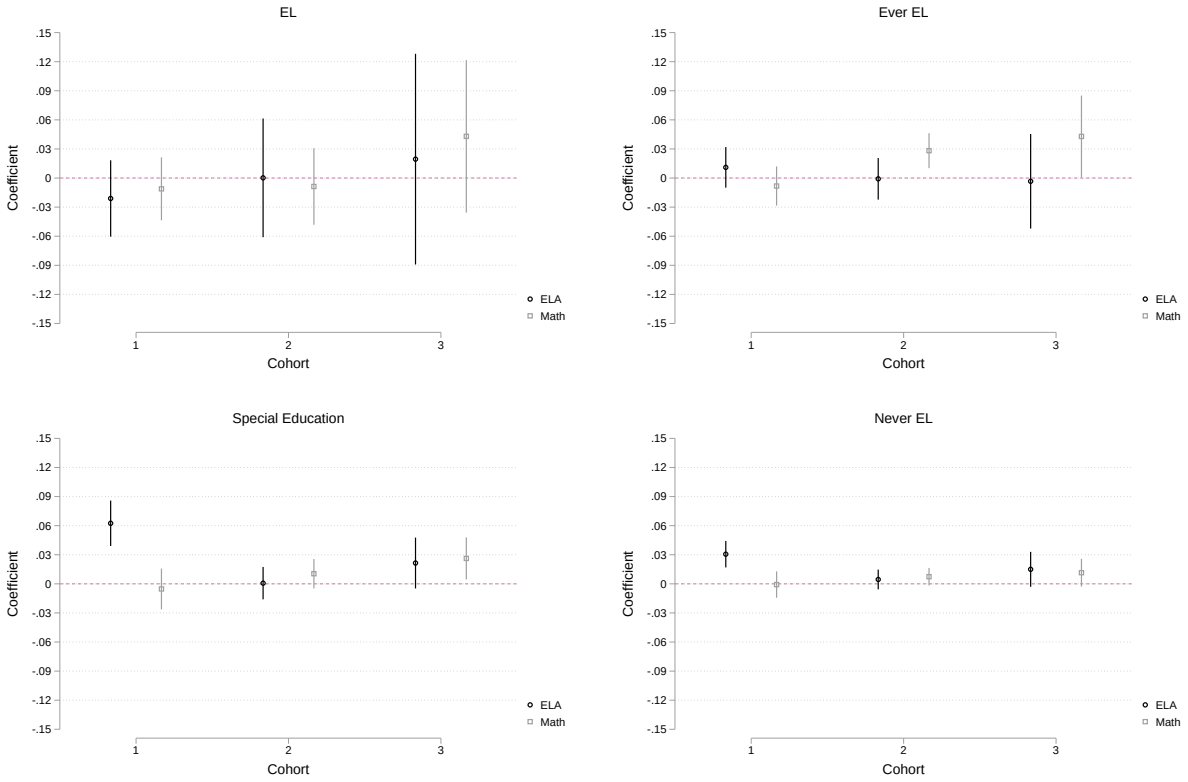


Figure (2) Effect by SEI Course Completion Year

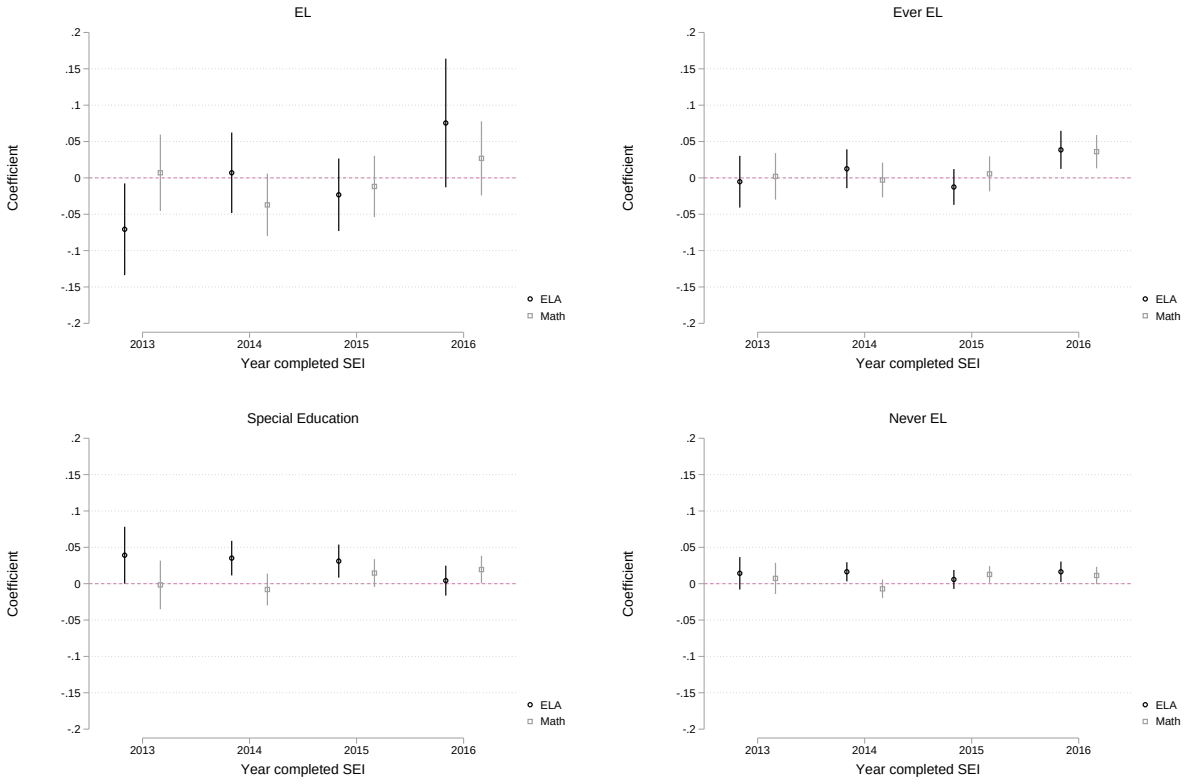


Figure (3) Effect by Grade Level

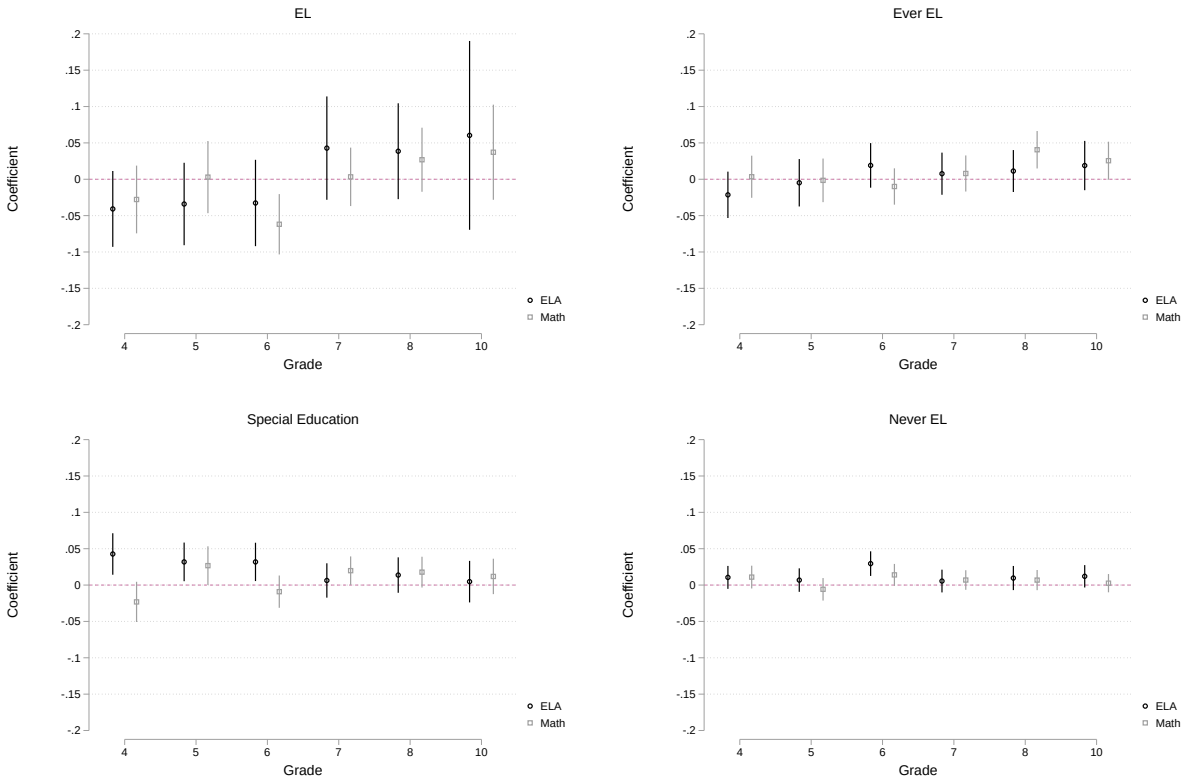


Figure (4) Effect by Prior MEPA/ACCESS Score

(a) ELA (b) Mathematics

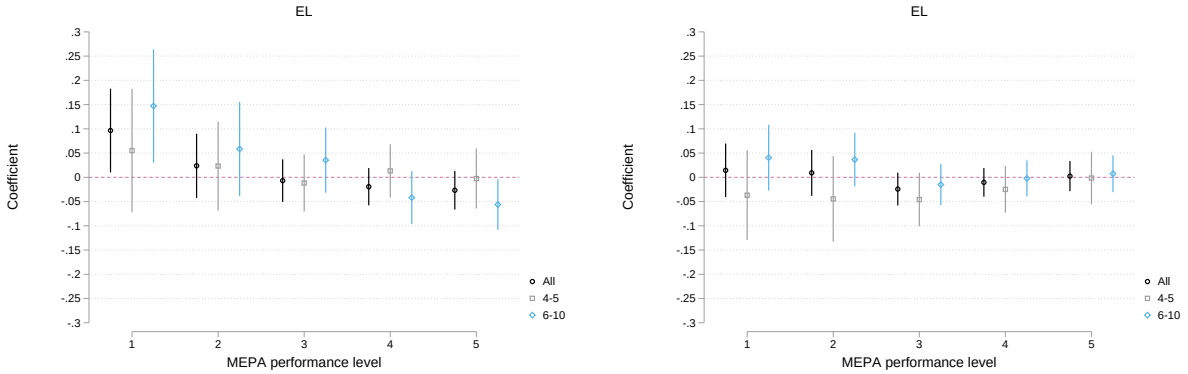


Figure (5) Effect by Teacher Years Since Hired Within District

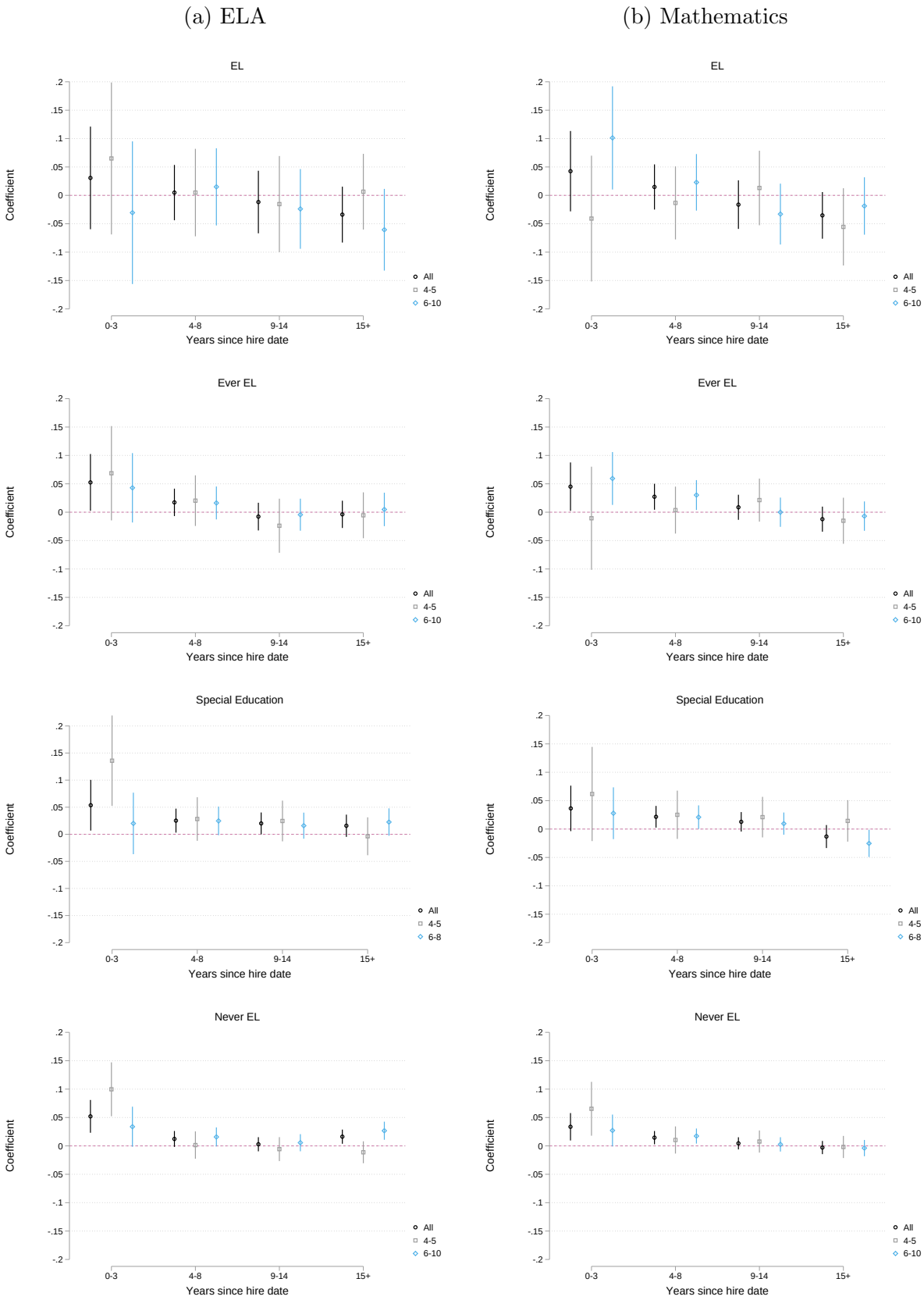
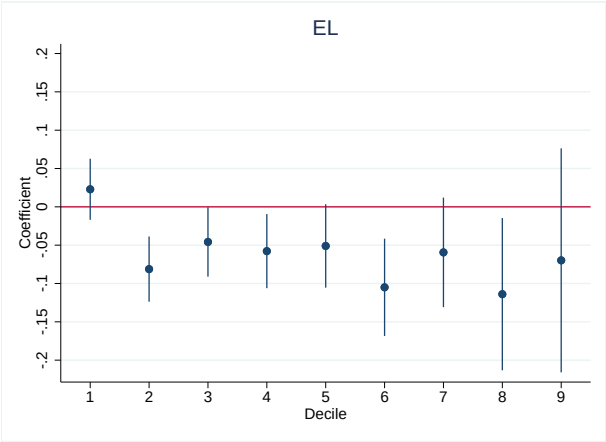
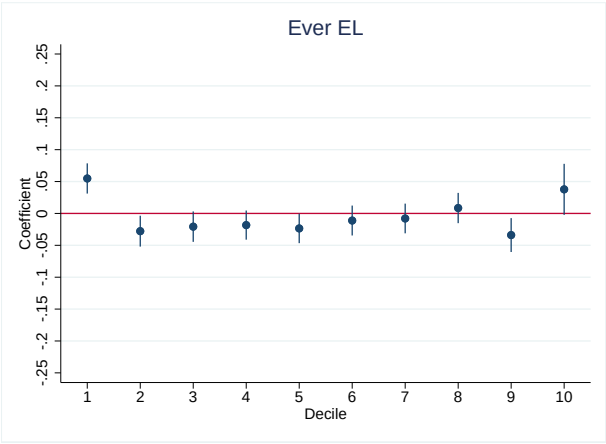
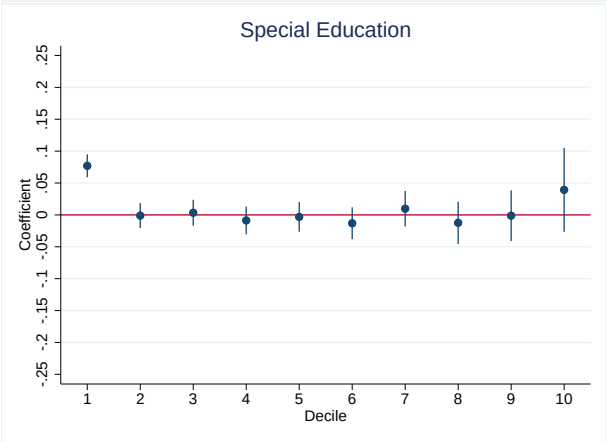


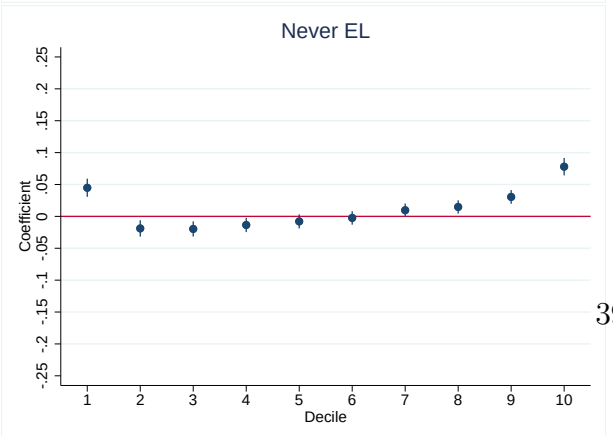
Figure (6) Effect by Prior Student ELA and Math Score

(a) ELA

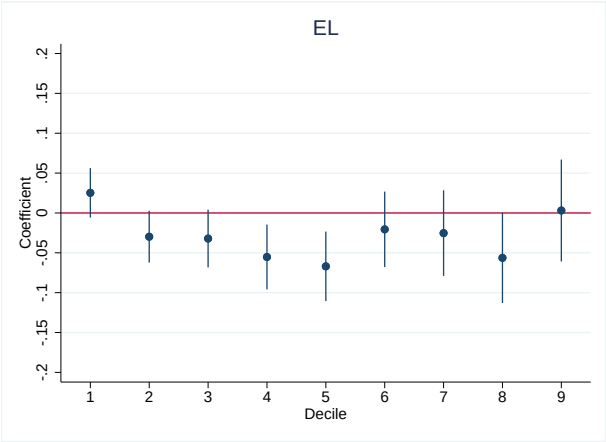


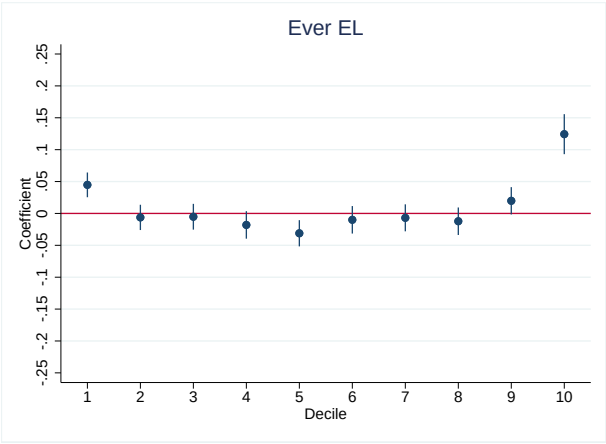


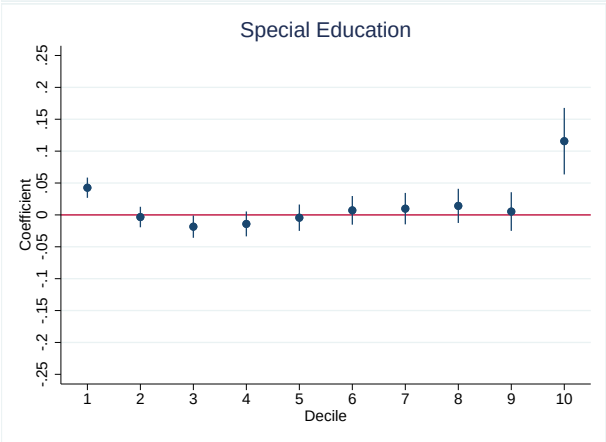




(b) Mathematics







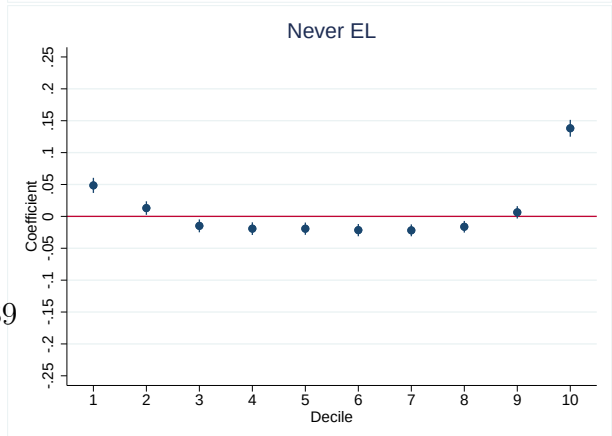
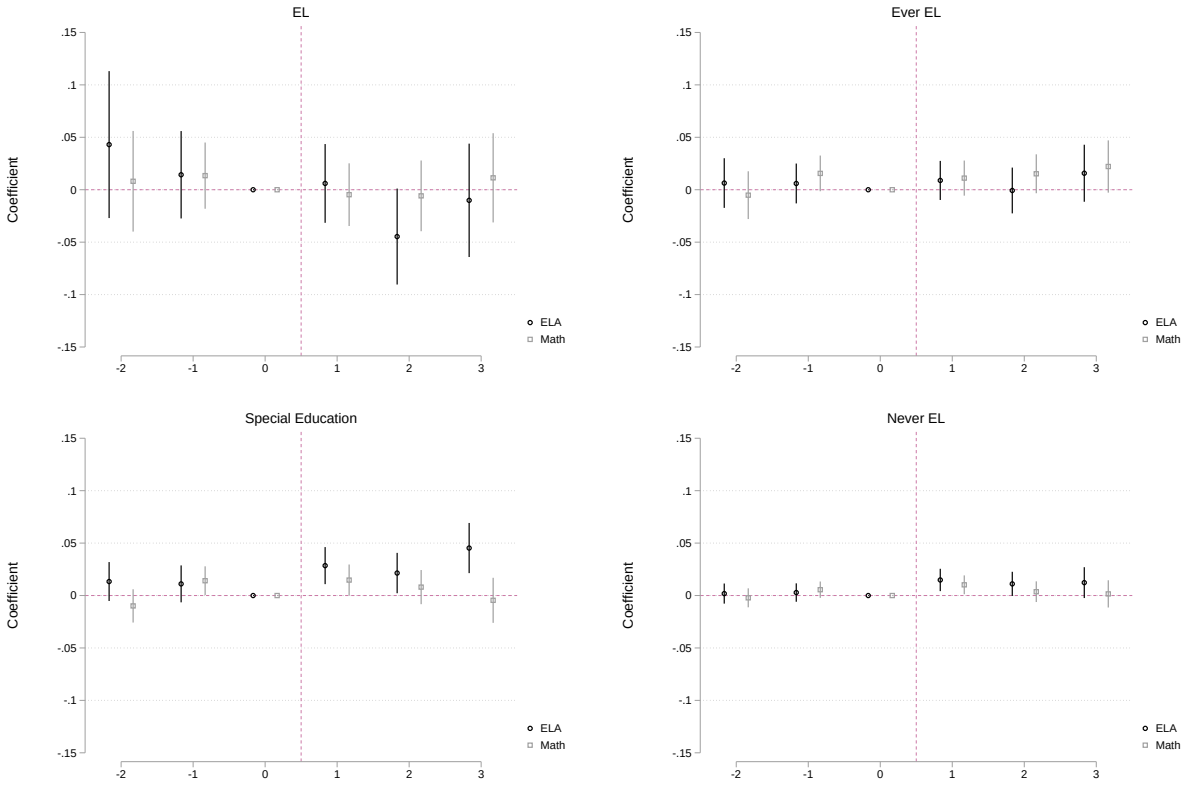


Figure (7) Identification Test



Tables

Table (1) Descriptive Statistics: ELA and Math Baseline Samples

	ELA				Mathematics			
	(1) EL	(2) Ever EL	(3) Special Ed	(4) Never EL	(5) EL	(6) Ever EL	(7) Special Ed	(8) Never EL
Student charecteristics								
Female	0.461	0.490	0.365	0.499	0.461	0.487	0.363	0.498
White	0.611	0.588	0.845	0.867	0.621	0.593	0.842	0.864
Hispanic	0.593	0.508	0.188	0.098	0.605	0.526	0.202	0.101
Black	0.246	0.198	0.143	0.102	0.257	0.205	0.146	0.105
Asian	0.150	0.213	0.037	0.058	0.127	0.197	0.037	0.058
Indian/Alaskan	0.021	0.031	0.016	0.012	0.028	0.037	0.017	0.012
Hawaiian	0.005	0.006	0.004	0.004	0.007	0.006	0.004	0.004
Free lunch	0.804	0.673	0.399	0.248	0.793	0.680	0.410	0.252
Reduced lunch	0.038	0.059	0.044	0.036	0.037	0.058	0.044	0.037
Grade 4	0.291	0.166	0.147	0.129	0.234	0.148	0.137	0.124
Grade 5	0.219	0.173	0.159	0.142	0.177	0.158	0.157	0.144
Grade 6	0.170	0.181	0.176	0.165	0.169	0.187	0.188	0.181
Grade 7	0.138	0.172	0.175	0.178	0.173	0.189	0.187	0.185
Grade 8	0.122	0.168	0.179	0.193	0.162	0.180	0.183	0.189
Grade 10	0.060	0.140	0.164	0.194	0.085	0.139	0.149	0.177
Student score performance								
Test score	-1.026	-0.326	-0.791	0.180	-0.905	-0.310	-0.783	0.153
Prior score in 1st dec.	0.470	0.212	0.331	0.074	0.418	0.213	0.340	0.084
Prior score in 2nd dec.	0.168	0.111	0.150	0.057	0.180	0.128	0.170	0.074
Prior score in 3rd dec.	0.116	0.098	0.115	0.063	0.180	0.128	0.170	0.074
Prior score in 4th dec.	0.086	0.101	0.100	0.075	0.078	0.085	0.087	0.075
Prior score in 5th dec.	0.063	0.096	0.086	0.089	0.063	0.084	0.073	0.084
Prior score in 6th dec.	0.041	0.098	0.074	0.110	0.048	0.080	0.062	0.095
Prior score in 7th dec.	0.031	0.099	0.064	0.137	0.041	0.083	0.055	0.112
Prior score in 8th dec.	0.016	0.088	0.047	0.161	0.031	0.084	0.047	0.134
Prior score in 9th dec.	0.008	0.071	0.027	0.167	0.024	0.101	0.042	0.194
Prior score in 10th dec.	0.002	0.027	0.007	0.067	0.005	0.042	0.011	0.076
Teacher characteristics								
Trained teacher	0.280	0.303	0.202	0.196	0.368	0.361	0.227	0.215
0-3 yrs experience	0.205	0.196	0.162	0.154	0.206	0.201	0.163	0.151
4-8 yrs experience	0.292	0.294	0.316	0.304	0.303	0.303	0.331	0.319
9-14 yrs experience	0.228	0.257	0.269	0.283	0.247	0.263	0.273	0.286
15+ yrs experience	0.274	0.252	0.253	0.259	0.244	0.233	0.233	0.243
Observations	55893	196200	281064	1572638	75908	235378	316814	1702007

Table (2) Average Effect of Training on Student Math and Reading Scores

	EL		Ever EL		Special Ed		Never EL	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				English Language Arts				
Trained	-0.014		0.005		0.022***		0.012***	
	(0.018)		(0.009)		(0.007)		(0.004)	
Trained x Full		0.015		0.018*		0.020**		0.013***
		(0.023)		(0.010)		(0.008)		(0.005)
Trained x Long Bridge		-0.022		0.007		0.024		-0.001
		(0.040)		(0.018)		(0.016)		(0.009)
Trained x Short Bridge		-0.039		-0.022		0.030*		0.020**
		(0.028)		(0.014)		(0.016)		(0.010)
Observations	55893	52007	196200	183963	281064	268294	1572638	1506210
Adjusted R^2	0.584	0.586	0.687	0.689	0.598	0.599	0.661	0.661
				Mathematics				
Trained	-0.007		0.011		0.009		0.006	
	(0.014)		(0.008)		(0.006)		(0.004)	
Trained x Full		0.016		0.022**		0.018**		0.010**
		(0.018)		(0.009)		(0.007)		(0.004)
Trained x Long Bridge		-0.006		0.007		-0.006		-0.012
		(0.027)		(0.013)		(0.015)		(0.009)
Trained x Short Bridge		-0.056**		-0.026*		-0.000		-0.008
		(0.024)		(0.016)		(0.015)		(0.010)
Observations	75908	69405	235378	216214	316814	299675	1702007	1617038
Adjusted R^2	0.625	0.625	0.750	0.752	0.657	0.658	0.748	0.748
Student Demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prior Math/ELA Score	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prior MEPA/ACCESS Score	Yes	Yes	Yes	Yes	No	No	No	No
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table (3) Effect on MEPA level (ELs only)

	All		Grades K-5		Grades 6-12	
	(1)	(2)	(3)	(4)	(5)	(6)
Trained	0.006		0.014		-0.011	
	(0.012)		(0.015)		(0.020)	
Trained x Level 1		-0.078***		-0.040		-0.141***
		(0.023)		(0.026)		(0.041)
Trained x Level 2		-0.060***		-0.034		-0.101***
		(0.018)		(0.022)		(0.031)
Trained x Level 3		-0.045***		-0.062***		-0.002
		(0.014)		(0.017)		(0.025)
Trained x Level 4		0.045***		0.057***		0.016
		(0.013)		(0.016)		(0.022)
Y-mean	0.512	0.512	0.486	0.486	0.551	0.551
Observations	65836	65836	39741	39741	26095	26095
Adjusted R^2	0.340	0.341	0.419	0.422	0.210	0.211
Student Demographics	Yes	Yes	Yes	Yes	Yes	Yes
Student Demographics x Grades K-5	Yes	Yes	NA	NA	NA	NA
Prior Math/ELA Score	No	No	No	No	No	No
Prior MEPA/ACCESS Level	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Teacher Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Teacher x Grades K-5 Fixed Effect	Yes	Yes	NA	NA	NA	NA

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Suggested Citation: Bruhn, J., Jones, N., Kanno, Y., & Winters, M.A. (2020). *Professional development at scale: The causal effect of obtaining an SEI endorsement under Massachusetts's RETELL initiative.* Wheelock Educational Policy Center. Available at wheelockpolicycenter.org.

OUR MISSION

The Wheelock Educational Policy Center (WEPC) conducts and disseminates rigorous, policy-relevant education research in partnership with local, state, and federal policymakers and stakeholders to improve pk-20 educational opportunities and holistic outcomes for underserved students.

www.wheelockpolicycenter.org
wheelockpolicy@bu.edu

Boston University Wheelock College of Education & Human Development
Wheelock Educational Policy Center

WHEELOCK EDUCATIONAL POLICY CENTER