ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-597

Does Teacher Professional Development Improve Student Learning? Evidence from Leading Educators' Fellowship Model

Ariana Audisio
Leading Educators

Rebecca Taylor-Perryman
Leading Educators

Tim Tasker
Leading Educators

Matthew P. Steinberg
Accelerate

Teachers are the most important school-specific factor in student learning. Yet, little evidence exists linking teacher professional development programs and the strategies or activities that comprise them to student achievement. In this paper, we examine a fellowship model for professional development designed and implemented by Leading Educators, a national nonprofit organization that aims to bridge research and practice to improve instructional quality and accelerate learning across school systems. During the 2015-16 and 2016-17 school years, Leading Educators conducted its fellowship program for two cohorts of instructional leaders, such as department chairs, mentor teachers, instructional coaches, and assistant principals, to provide these educators ongoing, collaborative, job-embedded professional development and to improve student achievement. Relying on quasi-experimental methods, we find that a school's participation in the fellowship program significantly increased student proficiency rates in English language arts and math on state achievement exams. The positive impact was concentrated in the first cohort and in just one of three regions, and approximately 80 percent of treated schools were charters. Student achievement benefitted from a more sustained duration of participation in the fellowship program, varied depending on the share of a school's educators who participated in the fellowship, and differed based on whether fellows independently selected into the program or were appointed to participate by their school leaders. Taken together, findings from this paper should inform professional learning organizations, schools, and policymakers on the design, implementation, and impact of educator professional development.

VERSION: May 2024

Suggested citation: Audisio, Ariana, Rebecca Taylor-Perryman, Tim Tasker, and Matthew P. Steinberg. (2024). Does Teacher Professional Development Improve Student Learning? Evidence from Leading Educators' Fellowship Model. (EdWorkingPaper: 22-597). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ah2f-z471

Does Teacher Professional Development Improve Student Learning?

Evidence from Leading Educators' Fellowship Model

Ariana Audisio
Rebecca Taylor-Perryman
Tim Tasker
Leading Educators

Matthew P. Steinberg
Accelerate

March 6, 2024

Forthcoming in the *Journal of Research on Educational Effectiveness*

The authors thank Leading Educators for providing the proprietary program data for this paper and Natalie Truong for research assistance. The research reported here was supported in part by the Carnegie Corporation of New York. Corresponding author: M. Steinberg (matthew@matthewpsteinberg.com).

Abstract

Teachers are the most important school-specific factor in student learning. Yet, little evidence exists linking teacher professional development programs and the strategies or activities that comprise them to student achievement. In this paper, we examine a fellowship model for professional development designed and implemented by Leading Educators, a national nonprofit organization that aims to bridge research and practice to improve instructional quality and accelerate learning across school systems. During the 2015-16 and 2016-17 school years, Leading Educators conducted its fellowship program for two cohorts of instructional leaders, such as department chairs, mentor teachers, instructional coaches, and assistant principals, to provide these educators ongoing, collaborative, job-embedded professional development and to improve student achievement. Relying on quasi-experimental methods, we find that a school's participation in the fellowship program significantly increased student proficiency rates in English language arts and math on state achievement exams. The positive impact was concentrated in the first cohort and in just one of three regions, and approximately 80 percent of treated schools were charters. Student achievement benefitted from a more sustained duration of participation in the fellowship program, varied depending on the share of a school's educators who participated in the fellowship, and differed based on whether fellows independently selected into the program or were appointed to participate by their school leaders. Taken together, findings from this paper should inform professional learning organizations, schools, and policymakers on the design, implementation, and impact of educator professional development.

Keywords: Teacher professional development; student achievement; professional learning communities; teacher coaching; distributed leadership

Introduction

After decades of research, there is consensus that teachers are the most important school-specific factor in student learning (Darling-Hammond, 2000; Nye et al., 2004; Wright et al., 1997). Students with more effective teachers not only score better on achievement tests (Aaronson et al., 2007; Rivkin et al., 2005; Rockoff, 2004), but they also graduate at higher rates (Koedel, 2008) and have higher post-secondary enrollment rates (Jackson, 2014). In the longer run, students of more effective teachers attain higher rates of post-secondary degrees (Lee, 2018), experience greater success in the labor market (Chetty et al., 2014), and generate significantly higher lifetime earnings (Chetty et al., 2014; Hanushek, 2011).

If student success depends to such a considerable degree on teacher effectiveness, then teacher effectiveness undoubtedly rests on the opportunities teachers have for developing strong instructional skills. To paraphrase a common saying among educators: good teachers are not born, they are developed. A growing body of research demonstrates that teacher effectiveness increases dramatically during their first three years in the profession (Boyd et al., 2005; Hanushek et al., 2005; Henry et al., 2011), with recent evidence suggesting that teacher effectiveness grows throughout the first decade of a teacher's tenure in the profession (Papay & Kraft, 2015). As any veteran teacher can attest, teachers improve not just in their first few years in the profession, but they are often required to continue to advance and grow their instructional practice over the course of their entire careers in response to a seemingly ever-evolving set of policies, standards, and expectations (Coolahan, 2002; Day, 1999). Teacher professional learning therefore represents the ongoing, career-long process of advancing one's practice, to which end teachers routinely engage with specific, time-limited initiatives or sets of activities (e.g.,

coaching) that together make up the distinct teacher professional development (PD) programs in which they may participate (Borko et al., 2010; Opfer & Pedder, 2011).

Despite the importance of educator professional development to student learning, we still know relatively little about the effectiveness of PD programs and the various strategies or activities that comprise them (Hill et al., 2021). Indeed, several scholars have observed that the role that individual teacher development programs (e.g., coaching) play in improving both teaching quality and student achievement is both undertheorized and understudied (Desimone, 2009; Wayne et al., 2008). Moreover, relatively few teacher PD programs have demonstrated consistent evidence of increased outcomes at scale, across both a broad range of settings and student groups, in part due to a lack of generalizable knowledge about the specific features of such programs that effectively improve student learning (Hill et al., 2021).

Where specific knowledge of the program features and their connection to student learning is available, challenges arise when considering how to assess impact when programs are differentiated or contextualized but maintain common elements. This challenge is only exacerbated by the relatively large degree of variation in program features and the multitude of educational contexts in which PD programs are implemented. All this programmatic and contextual variation limits the utility of most PD intervention studies for local educational agencies (LEA) and the professional learning providers supporting them. This is because these groups are concerned about successful implementation of a particular program within a specific, local context rather than a generic intervention implemented without concern for context (Hill et al., 2021).

In this paper, we aim to address these important gaps and understudied questions by examining both the features and the impact of an educator PD program implemented across a

small but diverse set of urban school districts in the United States. The PD program examined herein was designed and implemented by Leading Educators, a national nonprofit organization that aims to bridge research and practice to improve instructional quality and accelerate learning across school systems. During the 2015-16 and 2016-17 school years, Leading Educators ran a fellowship program for instructional leaders, such as department chairs, mentor teachers, instructional coaches, and assistant principals, within four different states that was designed to provide educators with ongoing, collaborative, job-embedded PD. Over the course of two school years, the program aimed to develop participating instructional leaders' (i.e., fellows') beliefs, knowledge, and skills to lead ongoing, pedagogical development for the teacher colleagues that they, in turn, supported in their schools. The learning these fellows led within their schools was specific to mathematics and English language arts (ELA) curriculum standards, and it aimed to shift the school's instructional culture, the fellows' own instructional practices, and the instructional practices of their teacher colleagues. The ultimate aim of all these adult learning efforts was to improve student learning in the schools supported by the program.

Our study addresses the following two questions: 1) Does the teacher fellowship model lead to improvements in student performance? 2) Do the effects of the teacher fellowship model vary by school characteristics or natural variation in program implementation (e.g., program duration; the proportion of leaders trained at each school; whether leaders enrolled as teams or individuals; and whether LEA leaders participated in the fellowship)? Relying on quasi-experimental methods for causal inference, including difference-in-differences and event study strategies, we find that a school's participation in the fellowship program increased student proficiency rates on both math and ELA state achievement exams. The impact of the fellowship model was concentrated in only one (Louisiana) of the three participating regions and only for

the first cohort of program participants. The impact of the fellowship model in Louisiana may be explained by several factors, such as the increased experience and time delivering the program among the program staff in the region, as well as the greater autonomy afforded to school leaders in this predominantly charter landscape. Notably, approximately 80 percent of the schools implementing the fellowship program are charter schools, though there was variation across the three program sites in the concentration of charter schools.

Leveraging variability in program implementation across multiple district contexts, our findings further suggest that student achievement benefitted more from a more sustained duration of teacher and leader participation in Leading Educators' fellowship program (i.e., two years rather than one year). We also show that specific program features generated differential benefits for student learning. Specifically, we show that student achievement varied depending on the share of a school's staff who participated in the fellowship model as instructional leaders and the extent to which those fellows independently selected into the program or were appointed to participate by their school leaders.

By conducting this rigorous impact evaluation of its educator fellowship model, Leading Educators acknowledges the importance of understanding the aspects of its service provision – and the contexts in which those services are provided – that may be most effective for improving educators' professional practice and student performance. Thus, this evaluation aims to inform other organizations on the design and implementation features related to improving educator PD and, ultimately, student learning. Further, we offer this paper as a 'call to action' for other professional learning partner organizations to intentionally and rigorously examine their programs' features and impact. Ultimately, findings from this paper should illuminate the

importance of transparency to understand the elements of program design and implementation that are most effective at improving instructional practice and student learning.

Related Literature

In a recent landmark study, Kraft and colleagues (2018) conducted a meta-analysis of 60 primary reports on teacher coaching, an increasingly prominent strategy employed by PD programs in education (Wei et al., 2009). Their meta-analysis only incorporated evidence from studies designed to establish causality (e.g., randomized controlled trials, quasi-experimental designs analyzed through a difference-in-differences approach). Their results showed that coaching significantly improves teachers' instructional practices and their students' learning outcomes. In their preliminary scan of the research literature, these researchers noted that several common programmatic features were associated with improved teacher practice and student achievement: an intense and sustained duration of teacher coaching and professional development; a focus on discrete instructional skills; and active learning. Their study also uncovered some important differences, however, corresponding to the type or focus of the coaching being provided. For instance, they found that content-specific coaching (e.g., coaching tailored to math instruction) to be marginally more effective at improving teachers' instructional practice and student achievement than was coaching focused more broadly on general pedagogical skills that are not tied to a specific content area. Furthermore, the pairing of coaching with other developmental strategies revealed some notable impact differences, with coaching being most effectively offered alongside group training or instructional resources (Kraft et al., 2018). Finally, their results revealed no significant differences according to the quantity of coaching received (Kraft et al., 2018), suggesting that coaching quality and the specific components – or combinations thereof – may be most important for its impact.

The impact of teacher coaching on student learning notwithstanding, the field lacks a robust body of evidence demonstrating that a more diverse range of strategies used in educator PD programs can consistently lead to significant improvements in student learning. For instance, in summarizing the findings of all 67 evaluations of programs funded by federal Investing in Innovation grants, Boulay et al. (2018) found that fully 57 of those offered PD as a key program component. The specific activities or strategies used by those PD programs included coaching, seminars or workshops, training on curriculum materials, and online or video-based resources. Of those 57 interventions, however, only 10 reported on an ELA or math student learning outcome, and only 6 (less than 10% of all evaluations) demonstrated evidence of at least one positive, statistically significant impact on student learning. Although the PD offered through these programs was found to incorporate an inconsistent set of strategies, it is worth noting five of these six were designed to provide content-specific training, a finding further supporting the importance of content-specificity in the design and implementation of educator PD.

Other research examining the optimal duration of educator PD programs, a primary implementation feature, offers up a contradictory set of findings. Another recent meta-analysis, this one incorporating evidence from 28 studies employing experimental or rigorous quasi-experimental designs, shows that whereas teacher PD programs improved student reading achievement overall, program duration did not explain any differences in the magnitude of the impact (Didion et al., 2020). Likewise, Lynch and colleagues (2019) found no association between the duration of PD and its impact on students' math and science achievement in their meta-analysis of 95 primary studies employing experimental and rigorous quasi-experimental designs. In light of these results, Kraft (2020) posits that PD programs will generally achieve smaller effects when they either require cumulative decisions or more sustained efforts over

time, suggesting that simply extending the period of implementation is no guarantee of greater program impact.

Other studies suggest there might be a threshold for the minimum amount of time required before the effects of an intervention with educators can be detected among their students. For example, in a study of many elementary schools, Timperley and Alton-Lee (2008) found that teachers needed two years to develop the skills to effectively lead inquiry cycles with their colleagues before those efforts produced significant improvements in student learning. Similarly, in a randomized controlled trial of the Teacher Potential Project program across 70 schools, Dolfin and colleagues (2019) found that the PD program, which incorporated teacher institutes along with coaching support for content, pedagogy, and professional learning culture, significantly improved students' test scores, but only after two years of implementation. In view of this mixed pattern of findings regarding program duration, Hill and colleagues (2021) argue that further exploration of the relationship between the duration of PD and the magnitude of its impacts on learning and achievement is necessary.

In contrast to the attention paid to program duration, the question of whether teachers should participate in PD programs alone or alongside school leaders (e.g., principals) or system leaders (e.g., principal managers, cross school instructional coaches, Directors of Literacy) has been paid remarkably little attention. One theoretical justification for involving both school and system (or district-level) leaders in professional learning rests on the importance of having a clear vision for instructional improvement and coherent structures to support it (Fullan & Quinn, 2015; Cobb et al., 2019). Indeed, Hill and colleagues (2018) link the null results of a math PD program to the lack of clear messaging and alignment from district leaders, offering some support for the value of school and system leader participation. In light of this, it is reasonable to

expect that involving school and system leaders in professional learning typically reserved for teachers could lead to additional benefits for student achievement.

A final, largely understudied aspect of teacher PD programs concerns the sustainability of their impact on longer-term changes, both to teachers' instructional practice as well as to their students' learning outcomes (Antoniou & Kyriakides, 2013; Desimone & Stuckey, 2014; Kennedy, 2016). Several school-based interventions have demonstrated positive, long-term impacts on non-academic outcomes like student-teacher relationships (Okonofua et al., 2022) or attendance (Elias et al., 1991), and there is some evidence that improvements in instructional practice and student outcomes can also be maintained by a school after the program or intervention that produced them has been discontinued. For example, Allen et al. (2011) found that teachers who participated in the My Teaching Partner-Secondary program made sustained changes to their instructional practice such that student achievement in the year following the intervention improved significantly but not the achievement of those students they taught while participating. Similarly, Timperley and Alton-Lee (2008) found some evidence that suggests student learning gains - and the PD-related processes that brought about those gains - were sustained for at least one year after their teachers stopped participating in a PD intervention. Yet, to our knowledge, no studies provide causal evidence that teachers' participation in a PD program can have a sustained, positive impact on their students' learning in subsequent years, while being taught by other teachers.

Related research has looked beyond the durability of specific program impacts to simply consider the lasting impact that teachers themselves have on their students' longer-term achievement. As an example, in a study comparing experimental and non-experimental approaches to estimating teacher effects on student achievement, Kane and Staiger (2008) find

that the influence an individual teacher has on student learning during any single school year is halved during each subsequent year. Specifically, teacher quality scores from the prior year explain roughly 50% of the variance in student achievement in the subsequent year, reduced to 25% of the variance in the year after that (Kane & Staiger, 2008). Taken together, these findings suggest that the achievement gains brought about by PD programs may indeed be sustained beyond the program's term. Gains in student learning are likely to degrade over time, however, particularly when students transition into the classrooms of teachers who may not have participated in the program that brought about those gains.

Leading Educators' Fellowship Program

Leading Educators enrolled instructional leaders into their fellowship program during the 2015-16 and 2016-17 school years. These fellows served in various roles at the school and occasionally district level, including grade-level chairs, department chairs, mentor teachers, instructional coaches, and assistant principals. Many, but not all, fellows were part or full-time classroom teachers as well as emerging leaders in their buildings. With the endorsement and support of their principals, fellows participated in leadership development sessions, coaching focused on instructional leadership, and leadership cycles of collaborative learning, bias awareness and mitigation training, and sessions designed to improve pedagogical and content knowledge. In turn, Leading Educators supported these fellows to take the knowledge and skills they had gained and translate them into cycles of collaborative learning that they would then facilitate for their teacher colleagues back at their schools each week. Below, we describe the programmatic features and implementation of the fellowship. We then provide evidence of changes in fellows' leadership competencies, mindsets, and pedagogical content knowledge, the proximal outcomes of interest in Leading Educators' fellowship program model.

Program Model

Leading Educators' program model was designed to improve instructional leadership, teaching, and student learning by incorporating best-practices around the design and content for the PD. Each of these best practices corresponds to the activities in the logic model (see Figure 1). First, the design exemplified the principle of distributed instructional leadership (e.g., Robinson et al., 2008; Seashore Louis et al., 2010), wherein responsibility for improving instruction lies not with a single individual, but rather is distributed across all the educators within a school and is only achieved through their collective efforts. Consistent with this principle, programming began with sessions for principals and fellows to design visions for distributed instructional leadership, in which fellows planned to lead teams of teachers to engage in shared collaborative learning to improve instruction across the school. Second, to support their work as instructional leaders, fellows attended workshops grounded in specific academic content areas (e.g., Cohen & Hill, 2001; Garet et al., 2016; National Academies of Sciences, Engineering, and Medicine, 2020) and the model provided fellows access to content expert coaches who would help them expand or amend their existing knowledge to improve their instructional leadership practice (e.g., Crandall, 1983; Timperley & Alton-Lee, 2008). Sessions also involved fellows learning how teachers' biases and expectations can shape the quality of the instruction they provide students, and in turn, the outcomes students are able to achieve (Gershenson et al., 2016; Pajares, 1992). Besides learning about their impact, fellows also learned specific strategies for interrupting such biases and low expectations in both themselves and others. Finally, fellows led collaborative learning (termed Cycles of Professional Learning) integrated directly into the school day in a job-embedded manner instead of being offered offsite or outside of normal school hours (e.g., Croft et al., 2010; Dana, 2010; Pacchiano et al., 2016)

following a cyclical approach wherein grade- and subject-area teams explored a single, consistent topic across multiple sessions (e.g., Crow & Hirsh, 2015; Wiener & Pimentel, 2017), with teams learning alongside one another, offering each other ideas and sharing feedback (e.g., Cohen, 2011; Lynch et al., 2019). In these Cycles of Professional Learning, fellows identified an area for instructional improvement, then designed and led sessions aligned to that topic approximately twice a month, totaling six to eight sessions per topic. In these sessions, teachers would learn together, apply new learning to instructional practice, and reflect on student work.

< Figure 1 about here>

Program Implementation

Districts or charter management organizations within four regions sponsored schools to participate in Leading Educators' fellowship program during the 2015-16 and 2016-17 school years: Louisiana (specifically the greater New Orleans and Baton Rouge areas); the greater Kansas City area; the Washington, DC metro area; and the greater Memphis area. Although Leading Educators began as a pilot program in New Orleans in 2008, Kansas City, Memphis, and Washington DC were identified as expansion sites based on potential interest, local funding opportunities, and student need.[1] Any public, open-enrollment district or charter school located within these four geographic areas was considered eligible to participate if (i) at least 70% of the student population qualified for free or reduced-price lunch, or (ii) the school was able to make the case that there was great need, such as student achievement that fell substantially below district averages. The program was heavily subsidized by a mix of local and national philanthropic dollars, but schools were also required to contribute an average of 30% of the cost, depending on the place-based grants in each location.

[1] We exclude Memphis from analyses because outcome data (i.e., student proficiency data) during the treatment year 2015-16 was unavailable.

Leading Educators recruited applicants, often teachers with leadership responsibilities (e.g., department chairs, mentor teachers) or other instructional leaders (e.g. instructional coaches, assistant principals) from schools that met those criteria through targeted invitations and open information sessions and events. Given the organization's expectation that partnering with school teams would have a deeper impact than working with individuals alone, potential candidates were encouraged to discuss participation in the fellowship program with their principal and other instructional leaders in their school and to apply as a team with a school-wide plan for distributed instructional leadership. If a school was unable to apply as a team, individuals were required to have their principal submit an online endorsement that described the leadership role and support the fellow would receive.

Identifying emerging leaders with a history of pedagogical effectiveness to enroll in the program was a key aspect of the program design because fellows were responsible for leading and developing the instructional practice of their teacher peers. The program designers believed that fellows with a strong foundation would most benefit from instructional leadership training. At this time, the Common Core shifts and standards were still relatively new, and most teachers had received little training on them. The program leaders believed fellows with a proven track record of success would be primed to internalize the new instructional shifts and train their peers. In contrast, providing whole-school training and coaching to all teachers would be much more expensive. Furthermore, the program designers theorized that reluctant or less effective teachers would be less likely to try out new instructional practices from an external PD provider than they would from someone in their school or district. Accordingly, supporting schools in retaining fellows by providing effective educators with additional career pathways was an important

potential mechanism through which the program could improve teacher instruction and student achievement.

To apply, prospective fellows were required to have instructional leadership responsibilities, at least two years of teaching experience, and to have received an effective or highly effective rating on their most recent teaching evaluation. All candidates completed the same three phase competency-based selection process, regardless of whether they applied as individuals or teams. The phases included a written application, instructional assessment, and interview and role play. Leading Educators' staff and consultants were required to calibrate on the formal admissions rubric before being eligible to score each of the three application phases. The admissions team identified a minimum cut score for each of these three phases, and candidates were either invited to the next phase or rejected based on this score threshold. For prospective fellows who applied as part of a team, the cut score was calculated across the team; however, members of a team could individually be denied if the assessors had significant concerns (see Table B5 in the appendix for additional details on the fellow selection process).

Once admitted to the program, fellows attended approximately two weeks of PD sessions during the summer and four workshops throughout the school year. In addition, fellows met monthly with a leadership coach for one-on-one or team-based support. When fellows enrolled as a team, the principal participated alongside them in a subset of sessions designed specifically to help principals support fellows' development. Over two years, this totaled 160 hours. Regarding coaching, fellows who enrolled as individuals were provided one-on-one coaching for about one to two hours per month during the school year (10 hours per year). Fellows who enrolled together as teams received the same number of coaching hours, but their coaching occurred in a group alongside the other fellows in their school. Fellows transferred their own

learning back to their schools through ongoing cycles of collaborative learning for teachers at their schools. Fellows often led cycles in both content areas, with 88% of fellows leading two or more cycles of 6-8 sessions. Each session of professional learning lasted 60-90 minutes; thus, fellows facilitated learning sessions for ELA and math teachers in their schools for approximately 12 to 24 hours per year. More of these cycles focused on ELA versus math, with 54% focusing on ELA, 34% on math, and the remaining 12% were judged to be focused on content-agnostic topics.

Leading Educators enrolled two cohorts of fellows from these four regions. For the first cohort, which began in 2015-16, much of their first year of programming focused on skills for coaching and leading teams. By the end of their first year in the fellowship, sessions transitioned to become content-specific and focused on the ELA and math shifts for college and career-readiness standards. The second cohort began in 2016-17 and included both new schools and new fellows from the first cohort of schools who joined in the second year. This group experienced two days of in-person, content-specific learning on the ELA and math shifts. The initial 2015-16 cohort experienced this same content through an online course. Later in the summer of 2016, both cohorts attended a regional institute where they prepared to design and lead ELA or math content-focused cycles of collaborative learning back at their schools. Throughout the following year, both cohorts participated in content-specific workshops focused on evidence of teacher and student learning aligned to the standards and shifts. However, due to unexpected program closures, the second cohort only received one year of programming, with both groups ending at the end of the school year in spring 2017.

Leading Educators administered an annual survey assessing improvement on the short-term outcomes of its logic model (see Figure 1): school culture that supports ongoing, content-

specific team learning, leadership competencies and equitable beliefs about students and instruction, and knowledge and practice alignment to rigorous, standards-based instruction. This annual diagnostic survey is included in Appendix B. Only fellows from Cohort 2016-17 participated in the diagnostic survey before and after beginning treatment in the spring of both 2016 and 2017, respectively. Fellows also submitted a performance assessment demonstrating their skills for leading adult learning as well as using student data. Together, these assessments suggested improvement was occurring for all short-term outcomes (see Figures 2 and 3). Although these results are limited by the sample of fellows responding to both surveys and represent only a subset of the classrooms included in the analytic sample included in this study, they do motivate and inform the current study. These results are further described below, to provide a detailed description of program implementation and its relationship to the proximal outcomes of interest: fellows' skills, knowledge, beliefs, and perceptions of culture.

Fellows reported modest growth in the instructional cultures within their schools as well as how frequently they exercised specific leadership competencies and their equitable beliefs about students (see Figure 2). For instance, at the individual level, fellows' beliefs about racially and ethnically diverse students, on average, became more equitable following one year of programming (see Figure 2). Data from the baseline annual diagnostic survey revealed that fellows were somewhat more advanced in their standards-aligned instructional knowledge and practices in mathematics compared to ELA. Specifically, fellows' average math score at baseline was 77% compared to 65% in ELA. Moreover, following one year of programming, the average score on the pedagogical knowledge and practice items increased by approximately 8 percentage points in ELA and math (see Figure 3).

< Figure 2 about here>

< Figure 3 about here>

In sum, findings from the extant research suggest that the duration of PD programs, the specific components or strategies they employ, and how those are combined together into a coherent program all warrant greater consideration. Data on the implementation of Leading Educators' program and the changes observed among the proximal outcome indicators described above provide additional rationale for examining the program's impact on student achievement. Accordingly, in the next section, we describe the data and empirical approach for estimating the impact of participation in the fellowship program on students' grade-level proficiency in ELA and math. By leveraging both program specific and publicly available data in the context of a differences-in-differences and event study approach, we examine the impact of Leading Educators' PD program on student achievement both during and after the fellowship.

Data & Sample

To estimate the impact of Leading Educators' PD program on student achievement, we construct a school-grade-year dataset for the 2009-10 through 2018-19 school years. The dataset contains grade-level data collected from the Education Data Portal at the Urban Institute (Urban Institute, 2022) and the U.S. Department of Education's Common Core of Data (CCD) for grades 3-8 in Louisiana, Washington D.C., and Kansas City.[2] The data contain detailed information at both the school-grade-year and school-year levels. At the school-grade-year level, data include: ELA and math proficiency midpoint from EDFacts, a Department of Education initiative, via the Education Data Portal of the Urban Institute, and student enrollment by ethnicity from the CCD. The ELA and math proficiency midpoint is the midpoint of the range

[2] We excluded special education schools, vocational education schools, alternative education schools and common core data reportable programs. Data from Memphis was excluded from the analyses because student proficiency data from one treatment year (2015-16) was unavailable.

used to report the share of students achieving "proficient" or "advanced" levels as defined by each State Education Agency on its English language arts (ELA) and math assessments. Ranges are used to protect the privacy of the students and are determined by the size of the group. For example, data from grade levels with fewer than 5 students are suppressed, and grade levels with the fewest students (6-15) are reported with the widest range (<50% or ≥50%). As the number of students increases, the magnitude of the range decreases, until there are more than 300 students in the group, at which point the exact proficiency is reported. At the school-year level, data include the estimated and modified estimated percentage of students living in poverty (MEPS) available beginning in the 2013-14 school year, collected from the Model Estimates of Poverty in Schools (MEPS) via the Education Data Portal of the Urban Institute. MEPS poverty is a statistical estimate of the percentage of school's students living in poverty and, according to EDFacts, is the preferred estimate for national-level analyses. MEPS modified poverty is a statistical estimate of the percentage of school's students living in poverty, modified to align with a measure of the school district's poverty and according to EDFacts this estimate is preferred for analysis of large school districts across time or across states.

To describe the characteristics of the treatment and comparison schools, we collected data from the CCD, school level indicators of magnet and charter status along with the number of full-time equivalent staff. We gathered the number of students in special education and English language learners (ELL) programs from the Civil Rights Data Collection (CRDC) via the Education Data Portal of the Urban Institute. Primary sourced data (i.e., data from Leading Educators) includes a list of schools enrolled in the fellowship program during school years 2015-16 and 2016-17, the number of fellows enrolled by school and program year, number of years enrolled, an indicator for schools that had one or more LEA leaders enrolled as fellows,

and an indicator that differentiates fellows who enrolled individually from fellows who enrolled as a school-based team. We excluded from the analysis file any schools or districts that received support from Leading Educators before the 2015-16 school year or during school years 2017-18 to 2018-19. Finally, we collected data about fellows' level of educational attainment, years of experience in K-12 education, and summary scores on the three components of the application process that participants completed prior to enrolling, which consisted of an online application, an interview day, and an instructional assessment that included a classroom observation and submission of student data. The application process varied slightly across regions and years.

Fellow Characteristics

Table 1 displays the share of fellows who entered with graduate degrees, their average years of experience in K-12 education, and the summary scores on the three components of the application process (i.e., written application, instructional assessment, and interview) on which participants were assessed prior to being accepted into the program. There were notable differences in the backgrounds of fellows who enrolled as individuals compared to those who enrolled as part of a school team. Compared to fellows enrolling as teams, those who participated as individuals were more likely to have a graduate degree, were less likely to identify as a person of color, scored slightly higher on each component of the selection process, and were more likely to complete the program.

Program completion was a challenge in all three regions. Withdrawal from the program was more common among fellows who enrolled with their team than for fellows enrolling as individuals. This may have been related to differences in recruitment. For instance, fellows who were part of a team may have been more likely to participate at the request of their principal, leading to more variability in their own personal interest and confidence in the program. Fellows

participating as individuals, on the other hand, may have learned about the program on their own and invested significant time in finding the funding required to attend and securing their principal's or supervisor's support. Additionally, as detailed above, some fellows received a shorter duration of programming due to program closure decisions made by the organization in spring 2017.

<Table 1 about here>

School Characteristics

Table 2 shows the baseline mean characteristics of the sample by treatment status for each of the two cohorts. It also shows the baseline differences between treatment and comparison group, after adjusting for district-grade fixed-effects on the grade-level characteristics and district fixed effects for school-level characteristics. For Cohort 1 schools that started the fellowship in 2015-16, data are based on school-grade-year analytic sample for ELA and math achievement outcome during the baseline (i.e., pre-treatment) year 2014-15, except for measures of English Learners and Special Education where baseline data was available in the 2013-14 school year. For Cohort 2 schools that started the fellowship in 2016-17, data are based on the year 2015-16. In the year prior to the start of the program for Cohort 1 schools, the absolute adjusted differences in both ELA and math proficiency across treatment and control schools were smaller than 0.05, satisfying What Works Clearinghouse (WWC) baseline equivalence standards. Cohort 1 adjusted differences on demographic characteristics were statistically significant on the shares of students who identified as Hispanic, White, and Asian and on the shares of students that were English Learners, received special education services, and were in grade 8. The baseline sample for Cohort 2 of schools satisfied WWC baseline equivalence standards for the math but not for the ELA outcome. The statistically smaller ELA baseline

proficiency for Cohort 2 compared to its comparison group may underestimate the size of the effect of the fellowship on ELA proficiency. Cohort 2 schools adjusted differences on demographic characteristics were statistically significant only on the shares of students who identified as Hispanic. The treatment sample for both cohorts had a larger proportion of charter schools but the adjusted differences were not statistically significant. In Washington, D.C, the targeting of charter schools was intentional, as Leading Educators was also co-designing a customized teacher leadership program in partnership with D.C. Public Schools that fell outside of the Fellowship model. Consequently, the D.C. Public Schools were not eligible for the program. In Louisiana, after Katrina and the formation of the Recovery School District, all schools in New Orleans became charter schools, and consequently predominantly charter schools were participants in the program in the Greater New Orleans region, though schools in surrounding regions that were not all charter also participated. These two regions contributed to the high percentage of charter schools in the program (see Table A1 for baseline characteristic by treatment status and region).

< Table 2 about here>

Program Characteristics

In the Leading Educators' Fellowship Program section above we described the vision, program components and focus of the fellowship model that were common in all schools and regions. Below (and in Figure 4) we describe the distribution of four elements of the program– saturation; enrollment type; duration; and local educational agency leader participation – where variation was observed at the fellow and school levels. We exploit this natural variation to explore heterogeneity in the impact of the fellowship program on student achievement and

inform future design of more cost-effective professional development. Below is a description of the four program elements and how their variation could moderate student outcomes.

Saturation. Saturation refers to the ratio of fellows in school-level roles to that school's overall number of full-time teaching staff. A higher saturation of fellows could lead to more coherence among the instructional leaders in a school and therefore higher achievement outcomes. At the same time, a smaller and more motivated group of fellows might benefit more from the fellowship model than a larger group, and too many instructional leaders in a building could be overwhelming or confusing for teachers. In addition, fellows who enrolled individually (resulting in low saturation) received individualized instead of group coaching, which may have better supported individual fellows in applying their learning to their school context. Figure 4 shows the distribution of saturation across school-grades. While the mean saturation across the two years was 0.06 (i.e., 6 percent of a school's full-time teachers were fellows), Figure 4 shows that a high proportion of schools had saturation levels of 0.02 to 0.03. **Enrollment type.** Enrollment type refers to whether fellows joined the program as individuals or as members of a school-based team. Individual enrollment involved fellows independently learning about the program, applying for funds, and securing their principal's endorsement. For team enrollment, principals or teachers learned about the program, principals secured funds for their schools, and then appointed a group of instructional leaders from their schools to join as fellows. Figure 4 shows that a similar number of schools had fellows who enrolled as individuals (14 schools) compared to those who enrolled as part of a school team (15 schools). While enrollment type and program saturation are related, we explore these two features separately as school teams also varied in size. Exploring heterogeneity in effects by both program features provides additional insight into the mechanisms by which variation in program implementation is related to variation

in program effects. **Duration.** Duration refers to the average amount of time fellows were enrolled in the program in each school. While the program was designed to have exactly one or two years of duration, depending on the region, some fellows withdrew before the end of the program and the program was terminated earlier than planned for some fellows. Using data about the exact enrollment and withdrawal days we created a continuous variable representing the amount of time in years and fractions of years that fellows were enrolled in the program. While the ideal amount of time for a PD program has important budgeting and practical implications, to our knowledge, this is the first study examining variation in student achievement with varying duration of PD. Figure 4 shows the distribution of program duration across school-grades in the fellowship schools. As expected, based on program design, the majority of the participants were enrolled between one and two years. Program duration of less than one year or between more than one and less than two years is the result of fellows withdrawing from the program, which could be correlated with selection into the program and could be a source of bias if the reasons for withdrawal were not random (e.g., due to unobserved fellow skills and attitudes). On average, fellows enrolled in the program for one year and three months.[3] **Local educational agency leader participation.** The fellowship program was designed to serve a variety of instructional leadership roles responsible for improving teaching and learning in schools. While most fellows were school-based, some district-level roles such as instructional coaches serving multiple schools also enrolled in the program. In addition to directly developing teachers they supported, LEA fellows enrolling alongside school-based fellows may have facilitated the creation of an aligned vision for instruction, an important system condition connected to gains in student achievement (Fullan & Quinn, 2015; Cobb et al., 2019). We defined the level of participation on

[3] See Figure A1 in the appendix for a distribution of duration by region.

the part of LEA leaders by two categories. The first category represents schools with fellows in LEA-level roles who enrolled alongside school-level fellows (there were seven schools in this category); and second, in most schools, only school-level fellows enrolled (there were 22 schools in this category).

<Figure 4 about here>

Empirical Approach

We rely on difference-in-differences and event study strategies to estimate the effect of Leading Educators' fellowship program implemented during the 2015-16 and 2016-17 school years. We also examine heterogeneity by program saturation and duration and the potentially differential effect for schools where LEA instructional leaders participated alongside school-based instructional leaders and for fellows enrolled as teams. We then explore heterogeneity by two school-level characteristics: the share of students in poverty and the share of students of color. Finally, we examine whether the impact of the fellowship model varied across region. We describe our main and heterogeneity models, below.

To address the first research question, we rely on a two-way fixed effects (TWFE) difference-in-differences approach to estimate the average effects of the fellowship program. This approach compares the change in ELA and math proficiency among schools participating in Leading Educators' fellowship program to changes in ELA and math proficiency among non-participating schools in the same district and grades in the years prior to the start of the program (2009-10 through 2014-15) and the years during and after the program (2015-16 through 2018-19). We also present results from a non-parametric event study model, which enables an assessment of the parallel-trends assumption underlying the difference-in-differences approach. In the context of the event study approach, the year-specific effects following the introduction of

the fellowship program provide insight into whether program effects accumulate or fade over time. Further, the year-specific effects prior to program implementation provide suggestive evidence on whether the TWFE approach meets the parallel-trends assumption; indeed, any statistically significant effects prior to the treatment period will signal pre-treatment differences in outcome trends, a violation of an important assumption of the parallel-trends assumption.

We specify the difference-in-differences model as follows:

$$Y_{gst} = \alpha + \gamma(Treat_{st}) + \theta_{gdt} + \delta_s + \varepsilon_{gst} \qquad (1)$$

where Y denotes the ELA and math proficiency midpoint for grade g in school s during school year t. The coefficient γ of $Treat$ represents the difference-in-differences estimate, where $Treat$ denotes whether school s received treatment in school year t (and all t years after treatment). The variables θ_{gdt} and δ_s denote district-grade-year fixed effects and school fixed effects, respectively, and ε_{gst} denotes the random error term. To account for group- and time-specific differences between the treatment and the comparison group, the school fixed effects account for all time-invariant (and unobservable) differences between schools, and the district-grade-year fixed effects account for all (unobserved) differences among grades in the same year and the same LEA (i.e., school district). The inclusion of district-grade-year fixed effects therefore restricts comparisons to schools in the same district and grade and school year; so, for charter school districts (i.e., charter management organizations with multiple charter schools), this means that we compare treated charter schools to untreated charter schools in the same charter management organization and ensures that we are not comparing outcomes in schools

with vastly different resources and governance structures.[4] We cluster the standard errors at the school level (to account for the correlation among students attending the same school).

We specify the non-parametric event study model as follows:

$$Y_{gst} = \sum_{j=-6}^{j=3} \gamma_j \left(Treat_{st+j}\right) + \theta_{gdt} + \delta_s + \varepsilon_{gst} \qquad (2)$$

where the γ_j coefficients of $Treat$ denote the *j*th pre- and post-treatment effects for students supported by teachers in schools with instructional leaders enrolled in the program. All other variables are defined as in Eq. (1), and we cluster the standard errors at the school level.

The fellowship program was implemented across two cohorts, the first starting in spring 2015 and the second starting in spring 2016. As recent literature suggests, the staggered timing of treatment may introduce bias by implicitly placing greater weight on units that experience the treatment for longer time periods as well as cases where treated units may be assigned negative weight due to the treatment-comparison cells in which they occupy (Baker et al., 2022; Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun & Abraham, 2021). To test the robustness of our TWFE and event study results to staggered treatment timing, we implement the procedure proposed by Sun and Abraham (2021) via the *sunab* command of the *fixest* package in R.

Exploring Heterogeneity by Program Implementation and School Characteristics

To address the second research question and to further understand the potentially differential influence of components of the fellowship program on student outcomes, we take advantage of natural variation among four program elements of the fellowship program model – saturation; enrollment type; duration; and local educational agency leader participation. To

[4] The majority of charters in the program were run by charter management organizations (CMO) with multiple schools in each LEA. There were 13 LEAs (district and CMOs) in the study that contained schools in both the treatment and comparison groups. Of these 13 LEAs, 4 were traditional public-school districts and 9 were CMOs.

estimate heterogeneity by program characteristics, we amend equations (1) and (2) by adding an interaction term for each program characteristic. The amended difference-in-differences and event study models are specified as follows:

$$Y_{gst} = \alpha + \gamma(Treat_{st}) + \beta(Treat_{st})(Char_s) + \theta_{gdt} + \delta_s + \varepsilon_{gst} \quad (3)$$

$$Y_{gst} = \sum_{j=-6}^{j=3} \gamma_j(Treat_{st+j}) + \sum_{j=0}^{j=3} \beta_j(Treat_{st+j})(Char_s) + \theta_{gdt} + \delta_s + \varepsilon_{gst} \quad (4)$$

where $Char$ denotes one of four program characteristics that vary in the post-treatment period among treated school *s*: (i) saturation; (ii) enrollment type (team or individual); (iii) duration; and (iv) local educational agency (LEA) leader participation. All other variables are defined as in equations (1) and (2) and we cluster the standard errors at the school level. *Saturation* is measured as the average proportion of instructional leaders (among the school's full-time teaching staff) in school *s* who participated in the fellowship program during the 2015-16 and 2016-17 school years (i.e., the proportion is the count of fellows in school *s* divided by the number of full-time equivalent teachers in school *s*); we calculate the heterogeneous effect by program saturation as a linear combination of $\hat{\gamma} + \hat{\beta}(Saturation_s)$. *Team* is an indicator that equals 1 if fellows in school *s* enrolled as a team and 0 if fellows enrolled as individuals; we calculate the heterogeneous effect by enrollment type as a linear combination of $\hat{\gamma} + \hat{\beta}(Team_s)$. *Duration* is a continuous measure of the average number of years (maximum 2 years) during which fellows were enrolled in the program and it was computed by adding the fractions of years of enrollment of all fellows in a school and dividing it by the total number of fellows during the two program years; we calculate the heterogeneous effect by program duration as a linear combination of $\hat{\gamma} + \hat{\beta}(Duration_s)$. *LEA Leader Participation*, or *TeachLEA*, is an indicator that equals 1 if both school-level fellows and LEA-level fellows in school *s* were enrolled in the

program and 0 if only school-level fellows enrolled in school *s*; we calculate the heterogeneous effect by whether LEA leaders participated as a linear combination of $\hat{\gamma} + \hat{\beta}$ when $TeachLEAs = 1$, and $\hat{\gamma}$ when $TeachLEAs = 0$. Finally, we examine potential heterogeneity in the effect of the fellowship program by two school characteristics: (i) proportion of students in poverty; and (ii) proportion of students of color. For each of these two school characteristics, we sort schools into terciles based on baseline (i.e., pre-treatment) data and separately re-estimate equations (1) and (2), by tercile. These heterogeneity analyses provide additional insight into the potentially differential effect of treatment for schools that vary (endogenously) on these program implementation and school-level characteristics, and do not provide evidence on the causal effect of program implementation and school-level characteristics on student achievement outcomes.

Results

Main Effects

Table 3 presents the difference-in-difference estimates of Leading Educators' fellowship program on ELA and math proficiency. On average, pooled TWFE difference-in-differences results for both math and ELA are positive though not statistically different from zero (results based on the Sun & Abraham (2021) approach which account for staggered treatment timing are also positive, of similar magnitude and are marginally statistically significant). Event study results provide no evidence of statistically significant pre-treatment effects in any of the six pre-treatment years (see Table 4 and Figure 5). In Figure 5 (and Table 4), the pre-treatment trends for math achievement approximate zero, while the pre-treatment trend in ELA achievement suggests a potential decline even as the pre-treatment effects individually are not statistically different from zero[5].

[5] We checked the robustness of the findings to removing each individual school and found that no individual schools were driving these results. Result tables are available upon request.

The event study estimates summarized in Table 4 reveal differential effects across the post-treatment period. For ELA, we find that student proficiency rates improved by 7.6 percentage points two years after the first year of treatment, with marginally significant improvements of 4.5 and 7.8 percentage points, respectively, during the first year and three years after the fellowship began (the effect 1 year after the fellowship began (5.4 percentage points) is qualitatively the same as the ELA effect during the first year of the fellowship). For math, the post-treatment effects show positive and statistically significant improvements in math proficiency of 6.6 percentage points one year after the first year of treatment and marginally significant improvements of 4.6 percentage points two years later. Given that content-specific educator professional development was not introduced into the fellowship program until the second year of the program (i.e., 2016-17 school year), these results showing that the greatest impact of the fellowship occurred two and three years after the first year of treatment support Leading Educators' expectation that content-specific instructional development leads to student learning gains. Notably, difference-in-differences and event study estimates are robust to models that account for staggered treatment timing (i.e. two cohorts of fellows) based on the procedure introduced by Sun & Abraham (2021). We also show that the average effect of the fellowship program was larger among Cohort 1 treatment schools than among Cohort 2 treatment schools (see Table 3). The larger effects, on average, among Cohort 1 schools suggests that the combination of one year of leadership programing plus one year of content-specific programing was likely more effective than just the one year of content-specific professional learning alone that Cohort 2 schools received.

< Table 3 about here>

< Table 4 about here>

< Figure 5 about here>

The event study results (see Table 4) indicate that ELA effects were larger in magnitude and more durable than math effects, which differs from many empirical studies of educational interventions that show stronger effects on math scores than reading scores (Hansen et al., 2018; Cronin et al., 2005). Leading Educators' internal program evaluation data suggest a similar increase in fellows' knowledge and practice scores in ELA and math; however, fellows' baseline math knowledge and practice were higher compared to ELA scores (see Figure 3). This could suggest that fellows may have had a stronger foundation in math and were therefore more likely able to immediately support the teachers in their school in improving math pedagogy. In contrast, fellows may have needed more time to improve ELA pedagogy, but changes may have been more drastic from prior practice and therefore produced a larger and more enduring effect. Another potential explanation comes from implementation data suggesting that fellows led more ELA content cycles compared to math content cycles, which would have resulted in more teachers improving their ELA classroom practices compared to math teachers.

To further contextualize the magnitude of the effects on student achievement, we convert the estimated change in the proportion of students proficient into: (i) percent change in the share of proficient students; and (ii) student-level change in standard deviation units. Relying on the difference-in-differences estimates from the TWFE approach, the ELA effect across the 4 post-treatment years represents a 6.1 percent (0.08 *SD*) increase in the proportion of students that are proficient, while the math effect represents a 6.2 percent (0.0.05 *SD*) increase.[6] The cohort 1

[6] We estimated the percent change in ELA (math) proficiency by dividing the effect expressed in the percentage point change in the proportion of students academically proficient by the weighted average proficiency rate at baseline (2014-15 and 2015-16) among the comparison schools. We estimated the effect expressed in student-level standard deviation units by dividing the effect expressed in the percentage point change in the proportion of students academically proficient by the weighted school-grade-year standard deviation at baseline (2014-15 and 2015-16) among all schools and then multiplying the result by the weighted average (2014-15 and 2015-16) of the proportion of students academically proficient (or advanced).

effect in ELA represents an 11.2 percent (0.14 *SD*) increase in the proportion of students that are proficient, while the math effect represents a 9.7 percent (0.08 *SD*) increase. The cohort 2 effect in ELA, although not statistically significant, represents an 8.5 percent (0.11 *SD*) increase in the proportion of students that are proficient, while the math effect represents a 5.8 percent (0.05 *SD*) increase.[7]

Effects by Program Saturation

Table 5 summarizes the effects at each of four levels of program saturation. Effects are larger at the lowest level of saturation with the largest effects occurring one and two years after the first year in ELA (0.17) and one year after in math (0.16). Effects are smaller in the second quartile when the ratio of fellows to full-time teachers is 0.04, with significant effects of 0.10 three years after for ELA and effects of 0.10 one year after for math. The effect becomes non-statistically significant when saturation increases to 0.07 for all years. Effects become significantly negative in the fourth quartile when the saturation mean is 0.11. The largest significant effects in the fourth quartile are one year after the first fellowship year for ELA (-0.15) and three years after in math (-0.19).

The negative results for higher levels of program saturation might seem counterintuitive but may be explained by the inefficiencies in the distribution of leadership roles, differences in program design and enrollment method. A large proportion of instructional leaders in a school building (among the school's full-time teaching staff) could cause inefficiencies in the distribution of leadership responsibilities and may translate into confusing messages for teachers.

[7] In terms of the event study estimates, the largest statistically significant effects expressed in percent change and student-level standard deviations are evident 2 years after the first year of treatment for ELA, representing a 13.7 percent (0.17 *SD*) increase, and 1 year after for math, representing a 14 percent (0.12 *SD*) increase. Two and three years after, effects ranged from significant to marginally significant in ELA and from non-significant to marginally significant in math. Magnitudes expressed as percent change in the proportion of students that are proficient or advanced ranged from 7.5 percent (0.07 *SD*) increase 3 years after in math to 14.1 percent (0.18 *SD*) increase 3 years after in ELA.

Additionally, the coaching time received by fellows in lower saturation schools was likely more personalized as individuals received one-on-one coaching whereas teams received group coaching. These findings may support the importance of differentiating programming supports according to baseline knowledge and experience, and the importance of building investment at all levels. Further, the teams and individual programs operating simultaneously differed in recruitment and enrollment method, which are both theoretically and empirically correlated with program saturation.[8] As discussed above, fellows who chose to enroll individually may be more motivated than fellows who were invited or required to participate by their principal. It may also be that the greater number of fellows a school leader chose to send reduces the likelihood that those fellows are prepared to engage in the program. Baseline skills, as measured by graduate degrees and scores from the selection process, were lower for team enrollment than for individuals. This may have led to a range in fellow readiness to adopt new practices and share them with their teacher colleagues. Another potential explanation is that with lower levels of skills, team-enrolled fellows were more negatively impacted by time taken away from instructional planning and preparation or other critical responsibilities. Given the lack of access to aligned curricular materials, fellows would have been pulled between creating or adapting their own materials, improving their practice, and leading their colleagues through cycles of collaborative learning. Findings from the heterogeneity analysis of enrollment type lend support to these hypotheses and are presented next.

< Table 5 about here>

Effects by Enrollment Type

[8] Pearson's correlation between program saturation and program type is 0.5 and statistically significant at the 1% level.

Table 6 shows the heterogeneous effects by the two types of program enrollment (individual and team). The results show that there are slightly stronger effects in math associated with the individual schools compared to the team schools (effects for ELA are nearly identical in the individual and team schools). In ELA, for schools with individually enrolled fellows, the effect is marginally significant (0.07) two and three years after the first fellowship year, whereas there is a statistically significant effect in ELA (0.08) for team-enrolled fellows two years after. In math, the effects two (0.09) years after are positive and significant for the individually-enrolled fellows, while there is no significant effect in math for team-enrolled fellows in any of the four post-treatment years. Enrollment type did not predict a significant difference in student ELA or Math performance. It could be that the hypothesized greater motivation of individually-enrolled fellows is compensated by the increased coherence gained when the school principal and a few other instructional leaders in the school are learning as a team.

<Table 6 about here>

Effects by Program Duration

Table 7 shows the heterogeneous effects for different durations of fellow enrollment in the program. Each duration level represents the approximate mean value at each of four quartiles of program duration. Effects when fellows enrolled for about 15 months or less are non-statistically significant. Notably, schools where fellows enrolled for the maximum of two years experienced positive and significant effects in all four years in ELA and one and three years after the first fellowship year in math. The larger effects among this group occurred two and three years after completion of the fellowship for both ELA (0.14) and for math (0.09). Though some of the shorter duration was due to attrition from the program and so likely reflected fellow motivation, some of the differences in program duration are also the product of program design

and outside factors. In Washington D.C., fellows enrolled for one year, and in Kansas City, the program was closed in 2017 in part due to continued funding challenges, and so fellows who enrolled there in 2016 could only receive a maximum of one year of PD. Schools where fellows participated for two years increased student proficiency by 11 percentage points more in ELA and 8 percentage points more in math compared to schools where fellows participated for only nine months during these same two years. These results suggest fellows benefited from ongoing, sustained professional development and from the combination of content-focused pedagogical and general leadership development.

<Table 7 about here>

Effects by LEA Leader Participation

Table 8 shows the heterogeneous effects for schools where fellows in LEA-level roles enrolled alongside school-level fellows compared to those schools where only fellows in school-based roles were enrolled. The results show that when LEA-level leaders enrolled alongside school-level leaders, the pooled treatment effects are positive and significantly larger (0.12) for ELA but not for math. The year-specific effects show significant effects across all four years in ELA and one year after the first year of fellowship in math when LEA leaders enroll alongside school fellows. These results support the idea that instructional alignment between schools and LEAs matters and could be key for the sustainability of program effects. We cannot separate LEA leader selection into the program from the program effects, but selecting and training a few highly effective LEA-level roles could be a cost-effective way to improve teaching and learning. More research is needed to understand the extent to which these outcomes could be achieved in larger LEAs that vary in terms of schools and student enrollment.

<Table 8 about here>

Notably, the heterogeneous results by program implementation are purely correlational. We cannot know for sure if the variation in student outcomes was caused by variation in these components or by other causes that are associated with these components. The extent to which these moderators are associated which each other could help generate better hypotheses about the direction and cause of these associations. All four program components are positively and significantly correlated with each other. That is, higher saturation is positively associated with longer duration, team enrollment and LEA leader participation; longer duration is positively associated with team enrollment and LEA participation; and team enrollment is positively associated with LEA participation. Since there is a negative correlation between saturation and student outcomes, the positive association with the other 3 program components provides some support to the study hypothesis that targeting a smaller group of school-based instructional leaders could improve student outcomes compared to a larger group. At the same time, the positive association between the other three program components does not provide additional support for their causal moderation of student outcomes. Conceptually, the association between longer program duration, team enrollment, LEA participation and improved student outcomes could be related to greater engagement in the fellowship initiative. An LEA who was more invested would be more likely to enroll leaders at the LEA level, support principals to participate, and to do so for the full program.

Effects by School Characteristics

Tables 9 and 10 show the heterogeneous effects for schools that differ on the share of students in poverty; Table 11 shows the heterogeneous effects for schools that differ on the share of students of color. Notably, while the heterogeneity sample (with 11,678 school-year-grade observations) differs from the main analytic sample (with 14,454 school-year-grade

observations) due to availability of baseline (i.e., pre-treatment) school poverty and race/ethnicity data, the difference-in-differences and event study estimates across these two samples are identical (see Table A6). The heterogeneity results indicate that, on average, student ELA and math achievement benefitted more from the fellowship program in schools serving (relatively) fewer students in poverty (Tables 9 and 10) and (relatively) fewer students of color (Table 11). These heterogenous results suggest that, among the treatment schools, those schools with higher concentrations of students in poverty likely required additional programmatic supports and training to realize comparable educational benefits of the fellowship program as their (relatively) more advantaged treatment school peers.

<Table 9 about here>

<Table 10 about here>

<Table 11 about here>

Effects by Region

To explore potential heterogeneity in program effects across regions, we estimated models by region and find that the effects are concentrated in Louisiana, with no discernible effects in Kansas City or Washington D.C. (see Tables 12 and 13, and Figure 6). There are several potential explanations for this pattern. First, Leading Educators formed in Louisiana three years prior to expanding to other regions, and consequently this longer history may have fostered a stronger reputation in the region that could have increased motivation for teachers to engage in the school-based learning required by the program. Additionally, given this longer history, the program staff in Louisiana had more experience in their roles, which may have made them more effective coaches or facilitators. It is unlikely that the difference is due to any of the program implementation components alone. For example, while the average duration was higher in

Louisiana (Figure A1), the mean duration in Louisiana (1.53 years) was similar to Kansas City (1.46 years). Saturation was also similar across all regions, and there is no clear pattern distinguishing Louisiana in enrollment type or participation at the LEA level. Though Louisiana is relatively unique as a predominantly charter school region, the greater school-based autonomy enjoyed by charter schools (relative to their traditional district school counterparts) may have better supported the implementation and impact of the fellowship's programmatic efforts. Yet, the Washington DC region also served exclusively charter schools, so the heterogeneity in program effects across regions is unlikely due to greater school-based autonomy in charter schools alone. Some combination of these factors may have contributed to the stronger outcomes in Louisiana.

<Table 12 about here>

<Table 13 about here>

Conclusion

School districts nationwide invest heavily in developing their educators, annually allocating 3-5 percent of their budgets to fund teacher PD initiatives (Kraft et al., 2018). For example, in the 2016-17 school year, expenditures on public elementary and secondary schools in the United States totaled $739 billion, or $14,439 per public school student (National Center for Education Statistics, U.S. Department of Education); thus, the annual cost of teacher PD to districts was $22-37 billion, or $433-722 per public school student. Yet, teacher PD programs have historically failed to improve teachers' instructional practice or student achievement (Kraft et al., 2018), and there is limited evidence on the specific features of PD programs that are most salient for improving student achievement across contexts (Hill et al., 2021). Not only have recent meta-analyses been unable to shed light on the optimal duration of PD (Didion et al.,

2020, Lynch et al., 2019), but the lack of clear information on the resources and factors necessary for effective implementation may account for a large part of the difficulty in scaling research-based strategies (Hollands et al., 2016). This paper adds to the literature by describing in detail the implementation of a PD program and by identifying characteristics that influence the effect on student achievement as well as the extent to which that effect is sustained over time.

This paper's findings are consistent with other rigorous evaluations of educator PD programs that have shown a positive impact on student learning through the development of instructional leaders (e.g., Mihaly et al., 2022; Gates et al., 2019). Results herein provide evidence that Leading Educators' fellowship program increased student achievement in ELA and math across different levels of support and program characteristics. We note that the impact of the fellowship model was concentrated in only one (Louisiana) of the three participating regions and only for the first cohort of program participants, and that approximately 80 percent of program schools in our study sample were charter schools. The main effects of the fellowship model range from a 6–8-point increase in the percentage of students who are proficient or advanced and are sustained (in ELA) for at least two years after the program ends. The absence of significant impacts of the program model in Kansas City, where the program targeted more traditional district schools, and in Washington D.C., where the program was newer to the region and the maximum duration was only one year, may provide important programmatic and implementation context to inform future replication efforts. We also find heterogenous effects of the fellowship model that are larger when the program targets a small proportion of fellows in each school to implement its PD intervention, when fellows remain in the program for two years compared to shorter durations, or when fellows with LEA-level roles participate alongside fellows with school-level roles. Additionally, as these effects are at the school level, these

findings suggest improvements in educator effectiveness were achieved across the school. To our knowledge, this study is among the first to provide evidence that a PD program targeting instructional leaders can have a sustained, positive impact on student learning in the years after the program ends. Moreover, as Chetty and colleagues (2014) suggest, improving teacher effectiveness may have long-term outcomes on students' college attendance and lifetime earnings even after the initial test score gains fade out. Future evidence from randomized evaluations will be valuable to confirm and provide additional support for these findings.

The effects were achieved with a relatively low-cost investment of approximately $75 per student per year, primarily covering personnel costs related to the design and delivery of learning sessions and for providing coaching.[9] To understand the relationship between investment and effectiveness, Kraft (2020) gathered per-pupil cost information for 68 education interventions to propose per-pupil cost benchmarks where less than $500 is considered low, between $500 and $4,000 is considered moderate, and greater than $4,000 is considered large. Kraft contrasts a higher-cost tutoring intervention (effect size 0.23 *SD* with an annual cost of more than $2,500 per student) with universal free breakfast (effect size 0.09 *SD* with a cost of $50-200 per student) and argues that the impact of universal free breakfast may be considered more impressive given the relatively low cost. Foster and colleagues (2013) also reviewed the cost effectiveness of

[9] Equipment and materials costs: The program used Google Drive and Canvas to manage documents and share program information. Google Drive products can be used free of cost, and Canvas memberships were 14.95 per participant per year. Total expenses added to about $430 per fellow. Personnel costs: Each region was staffed by two coaches, one program manager, and one program director. A central design and facilitation team of four created sessions, assessments, and tools. Participants were offered 160 hours of professional development and 20 hours of coaching over the two-year period, both within and outside of the school day. Participants were encouraged to engage in ongoing learning during planning time for 2-3 hours per month during the school day. Personnel costs totaled about $7,800 per fellow. Facilities cost: Most activities were hosted in school or district buildings, but summer events were typically held at hotel or other conference centers. Total cost was about $2,300 per fellow. Overhead expenses: About $2500 per fellow, this covered operating expenses and central staff for the nonprofit, such as office space and human resources. Cost paid by students or families: There was no charge to students or parents. Sources of funding: Fellowship program costs were covered by a mix of local and national philanthropy (roughly 70%) and direct fees from schools (approximately 30%).

teacher PD programs relative to other interventions, and reported similar ranges in cost, with one program producing a modest effect (0.03 *SD*) on math achievement at $44 per student and a master's degree program at $702 per student producing larger effects (0.22 *SD*) on math achievement. Leading Educators' fellowship program's approach to job-embedded, ongoing learning resulted in whole school effects at a low cost by targeting a subset of fellows per school, enabling cost savings compared to a direct to teacher or student approach. When compared with these frameworks, this program produced and, in some cases, sustained, medium to large effects at low cost.

While this study provides new evidence on the importance of assessing the readiness of schools and leaders to engage in an initiative for ongoing, school-based learning, there are remaining limitations that the analysis was unable to address. First, the study was unable to control for unobserved differences that led some fellows and schools to seek out the program. Additionally, while heterogeneity analyses provide insight into how the effect of treatment for schools varies (endogenously) across program implementation and school-level characteristics, they cannot provide evidence on the causal effect of program implementation and school-level characteristics on student achievement outcomes. Furthermore, the results of the event study by quartile of saturation suggested strong, enduring effects could be achieved by a small group of fellows, but a likely higher baseline knowledge and skills of these fellows may indicate these effects are not entirely due to the program. A better understanding of the role of baseline knowledge, skills, and conditions could support the development of differentiated approaches that address all schools' readiness for shifting teaching and learning. Additionally, future studies would benefit from a more detailed accounting of how instructional leaders spent their time, especially in schools with just one or two instructional leaders facilitating school-based learning.

As the heterogeneous effects of saturation level may in part be explained by the importance of fellow motivation in seeking out learning opportunities, this study also supports further investigation into how motivation influences the degree to which schools implement and sustain new practices. Lastly, it is worth noting that the treatment schools were majority charter schools, and it is possible charter schools may have been able to more quickly have success in the program; with increased autonomy, they may have had greater agency to make changes to instructional leader roles, responsibilities, and daily instruction. Consequently, results may not be generalizable to contexts where leaders have less autonomy to make changes to instruction and collaboration time, be they district or charter. However, the findings of studies of similar approaches in a sample where the majority of schools were traditional public schools demonstrate this model can achieve results outside of charter contexts (Mihaly et al., 2022). Furthermore, because treated schools who were standalone charters do not contribute to the impact estimates, we are unable to evaluate the impact of the program in that context and results may not be generalizable to charters that are not in a network with other charter schools, such as those in a charter management organization. It is possible standalone charter schools may have benefited more from the program due to increased autonomy or due to a larger need for external support. More research will be valuable in demonstrating results across contexts. Further understanding of these factors can help a variety of school systems and nonprofit partners make strategic design choices that maximize the value of PD initiatives for all schools in an effort to more widely distribute the benefits of rigorous, job-embedded educator PD.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95-135.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333* (6045), 1034-1037.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, *34*(2), 275-318.

Antoniou, P., & Kyriakides, L. (2013). A dynamic integrated approach to teacher professional development: Impact and sustainability of the effects on improving teacher behavior and student outcomes. *Teaching and Teacher Education*, *29*(1), 1-12.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7-39.

Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates?. *Journal of Financial Economics*, *144*(2), 370-395.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133-180.

Borko, H., Jacobs, J., & Koellner, K. (2010). Contemporary approaches to teacher professional development. In P. Peterson, E. Baker, & B. McGraw (Eds.) *International encyclopedia of education (3rd ed.) (pp.* 548-556). Oxford, UK: Academic Press.

Boulay, B., Goodson, B., Olsen, R., McCormick, R., Darrow, C., Frye, M., ... & Sarna, M. (2018). *The Investing in Innovation Fund: Summary of 67 evaluations. Final Report. NCEE 2018-4013.* Washington, DC: U.S. Department of Education.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). Explaining the short careers of high-achieving teachers in schools with low-performing students. *American Economic Review*, *95*(2), 166-171.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200-230.

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

Cobb, P., Jackson, K., Henrick, E., & Smith, T. M. (2020). *Systems for instructional improvement: Creating coherence from the classroom to the district office*. Cambridge, MA: Harvard Education Press.

Cohen, D. K. (2011). *Teaching and its predicaments.* Cambridge, MA: Harvard University Press.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works.* New Haven, CT: Yale University Press.

Coolahan, J. (2002). Teacher education and the teaching career in an era of lifelong learning. *OECD Education Working Papers, 2*(1), 1-39.

Crandall, D. P. (1983). The teacher's role in school improvement. *Educational Leadership, 41*(3), 6-9.

Croft, A., Coggshall, J. G., Dolan, M., & Powers, E. (2010). *Job-embedded professional development: What it is, who is responsible, and how to get it done well.* Washington, DC: National Comprehensive Center for Teacher Quality.

Cronin, J., Kingsbury, G. G., McCall, M. S., & Bowe, B. (2005). *The impact of the No Child Left Behind Act on student achievement and growth (2005 edition).* Portland, OR: Northwest Evaluation Association.

Crow, T., & Hirsh, S. (2015). Learning team cycle of continuous improvement. *Tools for Learning Schools, 19*(1), 1-7.

Dana, N. F. (2010). Teacher quality, job-embedded professional development, and school-university partnerships. *Teacher Education and Practice*, *23*(3), 321-325.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1-44.

Day, C. (1999). *Developing teachers: The challenges of lifelong learning*. London: Routledge.

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review, 110*(9), 2964-2996.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181-199.

Desimone, L. M., & Stuckey, D. (2014). Sustaining teacher professional development. In L. E. Martin, S. Kragler, D. J. Quatroche, & K. L. Bauserman (Eds.), *Handbook of professional development in education: Successful models and practices, PreK–12* (pp. 467–482). New York, NY: Guilford.

Didion, L., Toste, J. R., & Filderman, M. J. (2020). Teacher professional development and student reading achievement: A meta-analytic review of the effects. *Journal of Research on Educational Effectiveness, 13*(1), 29-66.

Dolfin, S., Richman, S., Choi, J., Streke, A., DeSaw, C., Demers, A., & Poznyak, D. (2019). *Evaluation of the Teacher Potential Project*. Washington, DC: Mathematica.

Elias, M. J., Gara, M. A., Schuyler, T. F., Branden-Muller, L. R., & Sayette, M. A. (1991). The promotion of social competence: Longitudinal study of a preventive school-based program. *American Journal of Orthopsychiatry*, *61*(3), 409–417.

Foster, J. M., Toma, E. F., & Troske, S. P. (2013). Does teacher professional development improve math and science outcomes and is it cost effective?. *Journal of Education Finance*, 255-275.

Fullan, M. (1995). The school as a learning organization: Distant dreams. *Theory into Practice*, *34*(4), 230-235.

Fullan, M., & Quinn, J. (2015). *Coherence: The right drivers in action for schools, districts, and systems*. Thousand Oaks, CA: Corwin Press.

Garet, M. S., Heppen, J. B., Walters, K., Smith, T. M., & Yang, R. (2016). *Does content-focused teacher professional development work? Findings from three Institute of Education Sciences Studies*. Washington, DC: U.S. Department of Education.

Gates, S. M., Baird, M. D., Doss, C. J., Hamilton, L. S., Opper, I. M., Master, B. K., ... & Zabar, M. A. (2019). *Preparing school leaders for success: Evaluation of New Leaders' Aspiring Principals Program, 2012-2017*. Santa Monica, CA: RAND Corporation.

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, *52*, 209-224.

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

Hansen, M., Levesque, E. M., Valant, J., & Quintero, D. (2018). *Brown Center report on American education: Trends in NAEP math, reading, and civics scores.* Washington, DC: Brookings Institution.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466-479.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (No. w11154). Cambridge, MA: National Bureau of Economic Research.

Hawley, W., Jordan, J. I. & Landa, M. (n.d.). *Common beliefs survey: Teaching racially and ethnically diverse students.* Montgomery, AL: Learning for Justice. Retrieved from www.learningforjustice.org on October 28, 2021.

Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers: Early-career teacher effectiveness and attrition. *Educational Researcher*, *40*(6), 271-280.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*(4), 372-400.

Hill, H. C., Corey, D. L., & Jacob, R. T. (2018). Dividing by zero: Exploring null results in a mathematics professional development program. *Teachers College Record*, *120*(6), 1-42.

Hill, H., Papay, J., Schwartz, N., Johnson, S., Freitag, E., Donohue, K., ... & Williamson-Zerwic, B. (2021). *A learning agenda for improving teacher professional learning at scale.* Providence, RI: Research Partnership for Professional Learning.

Hollands, F. M., Kieffer, M. J., Shand, R., Pan, Y., Cheng, H., & Levin, H. M. (2016). Cost-effectiveness analysis of early reading programs: A demonstration with recommendations for future research. *Journal of Research on Educational Effectiveness*, *9*(1), 30-53.

Jackson, C. K. (2014). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, *32*(4), 645-684.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). Cambridge, MA: National Bureau of Economic Research.

Kennedy, M. M. (2016). How does professional development improve teaching?. *Review of Educational Research*, *86*(4), 945-980.

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, *64*(3), 560-572.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

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

Lee, S. W. (2018). Pulling back the curtain: Revealing the cumulative importance of high-performing, highly qualified teachers on students' educational outcome. *Educational Evaluation and Policy Analysis*, *40*(3), 359-381.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms.* Washington, DC: National Center for Special Education Research.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293.

National Academies of Sciences, Engineering, and Medicine. (2020). *Changing expectations for the K–12 teacher workforce: Policies, preservice education, professional development, and the workplace.* Washington, DC: National Academies Press.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.

Mihaly, Kata, Isaac M. Opper, and Lucas Greer, The Impact and Implementation of the Chicago Collaborative Teacher Professional Development Program. Santa Monica, CA: RAND Corporation, 2022. https://www.rand.org/pubs/research_reports/RRA2047-1.html.

Okonofua, J. A., Goyer, J. P., Lindsay, C. A., Haugabrook, J., & Walton, G. M. (2022). A scalable empathic-mindset intervention reduces group disparities in school suspensions. *Science Advances*, *8*(12), 1-10.

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research, 81*(3), 376-407.

Pacchiano, D. M., Whalen, S. P., Horsley, H. L., & Parkinson, K. (2016). *Efficacy study of a professional development intervention to strengthen organizational conditions and effective teaching in early education settings.* Evanston, IL: Society for Research on Educational Effectiveness.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307-332.

Papay, J. P. & 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*(1), 105-119.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrics*, *73*(2), 417-458.

Robinson, V. M., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly*, *44*(5), 635-674.

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

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Seashore Louis, K., Dretzke, B., & Wahlstrom, K. (2010). How does leadership affect student achievement? Results from a national US survey. *School Effectiveness and School Improvement, 21*(3), 315-336.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175-199.

Timperley, H., & Alton-Lee, A. (2008). Reframing teacher professional learning: An alternative policy approach to strengthening valued outcomes for diverse learners. *Review of Research in Education*, *32*(1), 328-369.

TNTP. (2012). *The irreplaceables: Understanding the real retention crisis in America's urban schools.* New York, NY: Author.

Urban Institute (2022). *Education data portal*. [Data set]. Washington, DC: Author. Retrieved from https://educationdata.urban.org/data-explorer/explorer

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, *37*(8), 469-479.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Dallas, TX: National Staff Development Council.

Wiener, R., & Pimentel, S. (2017). Practice what you teach: Connecting curriculum & professional learning in schools. Washington, DC: *Aspen Institute*.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67.

Young, V. M., Schmidt, R., Wang, H., Cassidy, L., & Laguarda, K. (2017). *A comprehensive model of teacher induction: Implementation and impact on teachers and students.* Menlo Park, CA: SRI International.

Tables & Figures

Table 1. Characteristics of Program Fellows

	Overall		Louisiana		Washington D.C.		Kansas City	
	Individual	Team	Individual	Team	Individual	Team	Individual	Team
Completed Program	0.80	0.60	0.50	0.36	1.00	0.80	0.86	0.83
Years of Experience	6.50 (4.43)	6.43 (4.35)	5.67 (3.14)	4.75 (3.59)	7.29 (5.19)	5.89 (3.10)	6.43 (5.06)	10.67 (5.47)
Graduate Degree	0.70	0.46	0.33	0.44	0.86	0.10	0.86	1.00
Participants of Color	0.33	0.53	0.33	0.57	0.60	0.89	0.14	0.00
Application Score	0.64 (0.25)	0.53 (0.18)	0.62 (0.23)	0.53 (0.19)	0.56 (0.28)	0.55 (0.21)	0.82 (0.16)	0.50 (0.14)
Interview Score	0.67 (0.24)	0.64 (0.2)	0.53 (0.3)	0.56 (0.08)	0.71 (0.16)	0.72 (0.28)	0.83 (0.11)	0.70 (0.12)
Instructional Assessment Score	0.59 (0.23)	0.56 (0.18)	0.45 (0.24)	0.54 (0.11)	0.75 (0.22)	0.54 (0.26)	0.58 (0.03)	0.69 (0.11)
Fellows	20	30	6	14	7	10	7	6
Schools	17	12	6	4	6	5	5	3

Notes: Each cell reports mean (standard deviation), except for *Completed Program, Graduate Degree* and *Participants of Color*, which report proportions. Table is based on data collected by Leading Educators as part of the fellow application process (conducted during spring 2015 and 2016). Sample includes all 50 fellows in the final analytic sample of schools. *Individual* includes fellows who enrolled in the fellowship as an individual teacher; *Team* includes fellows in which the school principal enrolled multiple teachers as a school-based team. *Completed Program* is the proportion of fellows who completed the intended years of programming (one or two years) or remained enrolled until the program closed (as in Kansas City). *Years of Experience* is defined as the number of years teaching in K-12 education as fellows reported having upon applying to the program. *Graduate Degree* is the proportion of fellows who reported earning a masters or doctorate upon applying to the program. *Application Score*, *Interview Score*, and *Instructional Assessment Score*, which are continuous measures ranging from 0 to 1, are from the application process (see appendix Table B5 for details on the application, interview, and instructional assessment).

Table 2. Baseline Characteristics, by Treatment Status

	Cohort 1 (2015-16)				Cohort 2 (2016-17)			
	Treatment	Comparison	All	Adjusted Difference	Treatment	Comparison	All	Adjusted Difference
ELA proficiency	0.50 (0.31)	0.55 (0.24)	0.54 (0.24)	-0.04	0.34 (0.18)	0.56 (0.24)	0.55 (0.24)	-0.10***
Math proficiency	0.43 (0.27)	0.45 (0.24)	0.45 (0.24)	-0.04	0.29 (0.18)	0.48 (0.25)	0.47 (0.25)	-0.05
Hispanic	0.21 (0.22)	0.11 (0.15)	0.11 (0.16)	-0.11***	0.07 (0.14)	0.12 (0.16)	0.12 (0.16)	-0.06***
Black	0.53 (0.31)	0.60 (0.35)	0.60 (0.35)	0.06	0.79 (0.26)	0.58 (0.35)	0.58 (0.35)	0.08
White	0.21 (0.20)	0.24 (0.30)	0.24 (0.3)	0.07***	0.12 (0.21)	0.25 (0.30)	0.25 (0.30)	-0.03
Asian	0.01 (0.02)	0.02 (0.04)	0.02 (0.04)	-0.01***	0.00 (0.01)	0.02 (0.04)	0.02 (0.03)	-0.00
Poverty	0.26 (0.07)	0.26 (0.09)	0.26 (0.09)	-0.01	0.27 (0.08)	0.25 (0.09)	0.25 (0.09)	0.02
Modified Poverty	0.33 (0.11)	0.29 (0.12)	0.29 (0.12)	-0.01	0.33 (0.09)	0.32 (0.12)	0.32 (0.12)	0.02
English Learners	0.09 (0.17)	0.08 (0.13)	0.08 (0.13)	-0.09**	0.03 (0.04)	0.08 (0.12)	0.08 (0.12)	-0.05
Special Education	0.11 (0.06)	0.11 (0.05)	0.11 (0.05)	-0.03***	0.15 (0.05)	0.10 (0.06)	0.11 (0.06)	0.02
Student-Teacher Ratio	13.88 (2.85)	14.96 (3.45)	15.01 (3.42)	1.10	12.91 (2.73)	14.39 (4.66)	14.34 (4.62)	-0.29
Magnet	0.09	0.07	0.07	0.06	0	0.06	0.06	0.00
Charter	0.80	0.40	0.41	0.29	0.81	0.37	0.38	0.00
grade 3	0.27	0.21	0.22	0.09	0.21	0.21	0.21	0.06
grade 4	0.2	0.21	0.21	-0.00	0.21	0.21	0.21	0.05
grade 5	0.16	0.2	0.20	-0.04	0.21	0.21	0.21	-0.03
grade 6	0.16	0.14	0.14	0.02	0.17	0.14	0.14	-0.03
grade 7	0.14	0.12	0.12	0.05	0.1	0.12	0.12	-0.03
grade 8	0.07	0.11	0.11	-0.06***	0.1	0.11	0.11	-0.03
Schools	14	428	442	NA	12	416	428	NA
School-Grade Observations	44	1,498	1,542	NA	42	1,448	1,490	NA

Notes: Each cell reports mean (standard deviation) or the adjusted difference, except for grade-level, magnet, and charter, which reports proportions or adjusted difference. The adjusted differences for ELA and Math proficiency and race/ethnicity characteristics are estimated following the model: $X_{gs} = \beta(Treat_s) + \theta_{gd} + \varepsilon_{gs}$ where X denotes a baseline covariate for grade g in school s. The coefficient β of $Treat_s$ represents the adjusted difference, where $Treat$ denotes whether school s was ever a treatment school in the corresponding cohort g. The variable θ_{gd} denote district-grade fixed effects and ε_{gs} denotes the random error term. The adjusted differences for all other school-level variables are estimated following the model: $X_s = \beta(Treat_s) + \theta_d + \varepsilon_s$. For Cohort 1 schools that started the fellowship in 2015-16, data are based on school-grade-year analytic sample for ELA and Math achievement outcome during the baseline (i.e., pre-treatment) year 2014-15, except for measures of English Learners and Special Education where baseline data was available in the 2013-14 school year. For Cohort 2 schools that started the fellowship in 2016-17, data are based on year 2015-16. Baseline data were unavailable for 3 Cohort 1 treatment schools; therefore, their baseline characteristics do not appear in the table above. *ELA* and *Math Proficiency* is based on data reported by EDFacts (via the Urban Institute's Education Data Portal) and is measured as the midpoint of the range used to report the share of students achieving "proficient" or "advanced" levels as defined by each State Education Agency on the state's English language arts (ELA) and math assessments. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels. Across the two cohorts, 13 Local Education Agencies (LEAs) were represented; 4 traditional districts and 9 charter management organizations. On average, charter LEAs have 1.33 schools in the treatment group and 4.89 schools in the control group.

Table 3. Difference-in-Differences Estimates of Fellowship Program on Student Achievement

	TWFE ELA	S&A ELA	TWFE Math	S&A Math
	(1)	(2)	(3)	(4)
Panel A: Pooled Effect				
Treat	0.034 (.042)	0.057* (.031)	0.029 (.044)	0.039* (.023)
Panel B: Cohort Effect				
Cohort 1 (2015-16)		0.062* (.035)		0.045** (.020)
Cohort 2 (2016-17)		0.047 (.062)		0.027 (.052)
Observations	14,454	14,454	14,454	14,454
R^2	0.88950	0.88974	0.87137	0.87155
Within R^2	0.00076	0.00286	0.00043	0.00182

Notes: Each column (within a panel) represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The S&A results are based on the procedure introduced by Sun & Abraham (2021) to account for staggered treatment timing. Cohort 1 includes schools that first implemented the fellowship program in school year 2015-16; Cohort 2 includes schools that implemented the program in just the 2016-17 school year. All regressions include district-grade-year and school fixed effects. Coefficients are statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table 4. Event Study Estimates of Fellowship Program on Student Achievement

	Dynamic ELA (1)	S&A ELA (2)	Dynamic Math (3)	S&A Math (4)
6 years before	0.102	0.129	0.028	0.016
	(0.075)	(0.095)	(0.097)	(0.120)
5 years before	0.039	0.066	0.010	0.034
	(0.087)	(0.074)	(0.064)	(0.050)
4 years before	0.042	0.080	0.029	0.004
	(0.082)	(0.058)	(0.088)	(0.080)
3 years before	0.025	0.032	-0.006	-0.005
	(0.054)	(0.049)	(0.041)	(0.047)
2 years before	0.015	0.024	0.021	-0.008
	(0.037)	(0.034)	(0.031)	(0.041)
1st year of fellowship	0.045*	0.043	0.015	0.018
	(0.026)	(0.032)	(0.021)	(0.024)
1 year after	0.054	0.044	0.066**	0.060**
	(0.034)	(0.037)	(0.026)	(0.030)
2 years after	0.076**	0.073**	0.046*	0.038
	(0.033)	(0.032)	(0.023)	(0.033)
3 years after	0.078*	0.072	0.035	0.039
	(0.046)	(0.051)	(0.032)	(0.030)
Observations	14,454	14,454	14,454	14,454
R^2	0.88964	0.88974	0.87144	0.87155
Within R^2	0.00195	0.00286	0.00098	0.00182

Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. *1st year of fellowship* is the 2015-16 school year for Cohort 1 and the 2016-17 school year for Cohort 2. The S&A results are based on the procedure introduced by Sun & Abraham (2021) to account for staggered treatment timing. All regressions include district-grade-year and school fixed effects. Coefficients are statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table 5. Heterogeneous Effects of Fellowship Program, by Level of Saturation

	2% Saturation	4% Saturation	7% Saturation	11% Saturation	Difference in 2% to 11% Saturation
ELA					
Treat	0.14 (0.029)***	0.08 (0.023)**	-0.01 (0.027)	-0.13 (0.048)***	-0.27 (0.061)***
1st year of fellowship	0.15 (0.047)***	0.09 (0.049)	-0.01 (0.061)	-0.14 (0.009)	-0.29 (0.079)***
1 year after	0.17 (0.070)**	0.10 (0.062)	-0.01 (0.058)	-0.15 (0.070)*	-0.32 (0.077)***
2 years after	0.17 (0.064)**	0.12 (0.058)	0.28 (0.057)	-0.09 (0.068)	-0.26 (0.000)***
3 years after	0.14 (0.046)**	0.10 (0.056)**	0.05 (0.086)	-0.02 (0.135)	-0.16 (0.122)
Math					
Treat	0.13 (0.035)***	0.07 (0.018)***	-0.01 (0.032)	-0.13 (0.063)***	-0.25 (0.066)***
1st year of fellowship	0.11 (0.050)*	0.05 (0.031)	-0.03 (0.044)	-0.14 (0.078)	-0.25 (0.089)**
1 year after	0.16 (0.052)**	0.10 (0.043)*	0.02 (0.044)	-0.08 (0.078)	-0.24 (0.089)**
2 years after	0.14 (0.039)***	0.08 (0.035)**	-0.00 (0.046)	-0.12 (0.077)	-0.26 (0.086)***
3 years after	0.15 (0.037)***	0.07 (0.038)	-0.04 (0.056)	-0.19 (0.092)*	-0.34 (0.093)***

Notes: Each cell shows estimates (in proportions) for treated schools based on the linear combination of $\hat{\gamma} + \hat{\beta}(Saturation_s)$ for different values of saturation. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. *Saturation* is measured as the ratio of fellows in school-level roles to that school's overall number of full-time teaching staff. The mean (standard deviation) of saturation is 0.06 (0.04). See Table A2 for estimates of full specification of equations 3 and 4 upon which these estimates are based. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table 6. Heterogeneous Effects of Fellowship Program, by Enrollment Type

	Individual	Team	Difference in Team Individual
ELA			
Treat	0.03 (0.067)	0.04 (0.046)	0.01 (0.077)
1st year of fellowship	0.05 (0.046)	0.04 (0.054)	-0.00 (0.085)
1 year after	0.05 (0.038)	0.06 (0.070)	0.01 (0.092)
2 years after	0.07 (0.041)*	0.08 (0.036)**	0.02 (0.078)
3 years after	0.07 (0.050)	0.09 (0.073)	0.02 (0.085)
Math			
Treat	0.06 (0.068)	0.00 (0.046)	-0.05 (0.076)
1st year of fellowship	0.03 (0.045)	-0.00 (0.052)	-0.031 (0.086)
1 year after	0.08 (0.040)	0.06 (0.056)	-0.02 (0.078)
2 years after	0.09 (0.040)**	-0.01 (0.050)	-0.1 (0.070)
3 years after	0.06 (0.074)	0.00 (0.048)	-0.06 (0.11)

Notes: Each cell shows estimates (in proportions) for treated schools based on the linear combination of $\hat{\gamma} + \hat{\beta}(Team_s)$, where $Team = 1$ if fellows in school s enrolled as a team and 0 if fellows in school s enrolled as individuals. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. See Table A3 for estimates of full specification of equations 3 and 4 upon which these estimates are based. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table 7. Heterogeneous Effects of Fellowship Program, by Level of Duration

	9 months	1 year	15 months	2 years	Difference in 9 months to 2 years
ELA					
Treat	-0.01 (0.026)	0.01 (0.025)	0.03 (0.026)	0.10 (0.036)**	0.11 (0.031)***
1st year of fellowship	-0.00 (0.049)	0.02 (0.043)	0.04 (0.039)	0.11 (0.028)***	0.11 (0.033)***
1 year after	0.01 (0.057)	0.03 (0.050)	0.05 (0.046)	0.11 (0.043)**	0.11 (0.049)*
2 years after	0.03 (0.054)	0.05 (0.049)	0.07 (0.046)	0.14 (0.041)***	0.11 (0.037)***
3 years after	0.03 (0.060)	0.05 (0.056)	0.07 (0.054)	0.14 (0.050)**	0.11 (0.031)***
Math					
Treat	-0.00 (0.032)	0.01 (0.033)	0.03 (0.036)	0.07 (0.051)	0.08 (0.038)*
1st year of fellowship	-0.02 (0.036)	-0.00 (0.030)	0.01 (0.026)	0.06 (0.026)*	0.08 (0.035)*
1 year after	0.03 (0.037)	0.05 (0.034)	0.06 (0.033)	0.11 (0.034)***	0.08 (0.030)**
2 years after	0.03 (0.041)	0.04 (0.034)	0.04 (0.030)	0.07 (0.036)	0.05 (0.053)
3 years after	-0.02 (0.042)	0.00 (0.035)	0.03 (0.032)	0.09 (0.036)**	0.11 (0.049)*

Notes: Each cell shows estimates (in proportions) for treated schools based on the linear combination of $\hat{\gamma} + \hat{\beta}(TimeEnrolled_s)$ for different values of enrollment time (i.e., duration in the fellowship). The mean (standard deviation) of time enrolled is 1.22 (0.46) years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. See Table A4 for estimates of full specification of equations 3 and 4 upon which these estimates are based. Coefficients are statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table 8. Heterogeneous Effects of Fellowship Program, by LEA Leader Participation

	School Fellows & LEA	School Fellows only	Difference in School Fellows & LEA and School Fellows only
ELA			
Treat	0.07 (0.044)	-0.05 (0.030)	0.12 (0.053)*
1st year of fellowship	0.08 (0.027)***	-0.03 (0.057)	0.12 (0.060)
1 year after	0.10 (0.044)*	-0.04 (0.059)	0.14 (0.067)
2 years after	0.12 (0.037)***	-0.01 (0.059)	0.13 (0.057)*
3 years after	0.11 (0.049)*	0.00 (0.067)	0.11 (0.062)
Math			
Treat	0.05 (0.055)	-0.02 (0.042)	0.07 (0.069)
1st year of fellowship	0.04 (0.032)	-0.04 (0.051)	0.08 (0.071)
1 year after	0.09 (0.036)**	0.02 (0.053)	0.06 (0.069)
2 years after	0.06 (0.032)	0.01 (0.059)	0.04 (0.074)
3 years after	0.07 (0.034)	-0.03 (0.082)	0.09 (0.098)

Notes: Each cell shows estimates (in proportions) for treated schools based on the linear combination of $\hat{\gamma} + \hat{\beta}(TeachLEA)$, where $TeachLEA = 1$ if both school-level fellows and LEA-level fellows were enrolled in the program in school s and 0 if only school-level fellows were enrolled in the program in school s. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. See Table A5 for estimates of full specification of equations 3 and 4 upon which these estimates are based. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) level.

Table 9. Heterogeneous Effects on ELA Achievement, by Tercile of School Poverty

	MEPS Poverty						MEPS Modified Poverty					
	Pooled T1 (1)	Dynamic T1 (2)	Pooled T2 (3)	Dynamic T2 (4)	Pooled T3 (5)	Dynamic T3 (6)	Pooled T1 (7)	Dynamic T1 (8)	Pooled T2 (9)	Dynamic T2 (10)	Pooled T3 (11)	Dynamic T3 (12)
Treat	0.108**		-0.017***		0.061		0.100**		0.004		0.050	
	(0.042)		(0.006)		(0.052)		(0.040)		(0.060)		(0.068)	
6 years before		-0.104**						-0.091*				
		(0.050)						(0.055)				
5 years before		-0.203**		0.092***		0.092		-0.202*		0.075		0.120
		(0.080)		(0.027)		(0.111)		(0.110)		(0.054)		(0.128)
4 years before		-0.131		0.192***		0.039		-0.113		0.142*		0.036
		(0.126)		(0.027)		(0.054)		(0.112)		(0.076)		(0.070)
3 years before		-0.089*		-0.075***		0.079		-0.084		0.006		0.094
		(0.048)		(0.027)		(0.081)		(0.054)		(0.051)		(0.087)
2 years before		-0.067		0.077		0.040		-0.077		0.019		0.060**
		(0.049)		(0.047)		(0.038)		(0.050)		(0.034)		(0.024)
1st year of fellowship		0.021		-0.017		0.086***		0.029		0.058		0.076***
		(0.049)		(0.053)		(0.024)		(0.048)		(0.054)		(0.018)
1 year after		-0.012		0.070*		0.098**		-0.013		0.082**		0.090
		(0.041)		(0.039)		(0.043)		(0.046)		(0.036)		(0.062)
2 years after		0.016		0.049***		0.138***		0.010		0.078		0.118***
		(0.043)		(0.016)		(0.018)		(0.051)		(0.052)		(0.021)
3 years after		-0.004		0.073***		0.140**		-0.005		0.037**		0.187***
		(0.047)		(0.007)		(0.068)		(0.052)		(0.018)		(0.038)
Mean (SD) poverty measure	0.17(0.06)	0.17(0.06)	0.28(0.02)	0.28(0.02)	0.35(0.02)	0.35(0.02)	0.17(0.07)	0.17(0.07)	0.31(0.02)	0.31(0.02)	0.42(0.06)	0.42(0.06)
Observations	3,797	3,797	3,929	3,929	3,952	3,952	3,838	3,838	4,027	4,027	3,813	3,813
R^2	0.89111	0.89255	0.88819	0.88911	0.81767	0.81819	0.90162	0.90279	0.88758	0.88815	0.86613	0.86684
Within R^2	0.01184	0.02488	0.00013	0.00831	0.00273	0.00559	0.00956	0.02138	1.42×10^{-5}	0.00506	0.00146	0.00679

Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. *MEPS Poverty* is a statistical estimate of the percentage of school's students living in poverty and according to EdFacts is the preferred estimate for national-level analyses. *MEPS Modified Poverty* is a statistical estimate of the percentage of school's students living in poverty, modified to align with a measure of the school district's poverty and according to EdFacts this estimate is preferred for analysis of large school districts across time or across states. The poverty terciles were created by calculating the school level mean poverty (modified poverty) from 2013-14 and 2014-15, where baseline poverty data was available. Each tercile includes 131 schools. All regressions include district-grade-year and school fixed effects. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table 10. Heterogeneous Effects on Math Achievement, by Tercile of School Poverty

Subject			MEPS Poverty				MEPS Modified Poverty					
Model	Pooled T1 (1)	Dynamic T1 (2)	Pooled T2 (3)	Dynamic T2 (4)	Pooled T3 (5)	Dynamic T3 (6)	Pooled T1 (7)	Dynamic T1 (8)	Pooled T2 (9)	Dynamic T2 (10)	Pooled T3 (11)	Dynamic T3 (12)
Treat	0.119***		0.023***		0.033		0.128***		0.001		-0.004	
	(0.038)		(0.005)		(0.066)		(0.039)		(0.057)		(0.065)	
6 years before		-0.113***						-0.149***				
		(0.034)						(0.049)				
5 years before		-0.115		0.045		0.026		-0.127		0.048		0.068
		(0.078)		(0.035)		(0.073)		(0.084)		(0.099)		(0.058)
4 years before		-0.201*		0.066*		0.138**		-0.209**		0.132***		0.128
		(0.116)		(0.035)		(0.068)		(0.090)		(0.035)		(0.093)
3 years before		0.022		-0.100***		0.064**		-0.027		-0.072*		0.066*
		(0.032)		(0.035)		(0.032)		(0.058)		(0.040)		(0.035)
2 years before		0.023		0.021		0.099***		-0.003		-0.021		0.108***
		(0.023)		(0.058)		(0.034)		(0.057)		(0.046)		(0.030)
1st year of fellowship		0.031		-0.009		0.023		0.036		0.020		0.008
		(0.034)		(0.062)		(0.036)		(0.035)		(0.070)		(0.035)
1 year after		0.026		0.127*		0.119***		0.032		0.088		0.097**
		(0.030)		(0.065)		(0.041)		(0.035)		(0.062)		(0.048)
2 years after		0.070**		0.064*		0.092*		0.067**		0.045		0.061
		(0.031)		(0.033)		(0.055)		(0.032)		(0.030)		(0.060)
3 years after		0.091**		-0.040***		0.115**		0.087***		-0.009		0.095*
		(0.037)		(0.007)		(0.050)		(0.032)		(0.037)		(0.055)
Mean (SD) poverty measure	0.17(0.06)	0.17(0.06)	0.28(0.02)	0.28(0.02)	0.35(0.02)	0.35(0.02)	0.17(0.07)	0.17(0.07)	0.31(0.02)	0.31(0.02)	0.42(0.06)	0.42(0.06)
Observations	3,797	3,797	3,929	3,929	3,952	3,952	3,838	3,838	4,027	4,027	3,813	3,813
R^2	0.88628	0.88779	0.85980	0.86039	0.81980	0.82074	0.89728	0.89847	0.85783	0.85871	0.84786	0.84834
Within R^2	0.01054	0.02361	0.00015	0.00431	0.00069	0.00589	0.01115	0.02258	9.95×10^{-7}	0.00620	8.29×10^{-6}	0.00320

Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. *MEPS Poverty* is a statistical estimate of the percentage of school's students living in poverty and according to EdFacts is the preferred estimate for national-level analyses. *MEPS Modified Poverty* is a statistical estimate of the percentage of school's students living in poverty, modified to align with a measure of the school district's poverty and according to EdFacts this estimate is preferred for analysis of large school districts across time or across states. The poverty terciles were created by calculating the school level mean poverty (modified poverty) from 2013-14 and 2014-15, where baseline poverty data was available. Each tercile includes 131 schools. All regressions include district-grade-year and school fixed effects. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table 11. Heterogeneous Effects on ELA and Math Achievement, by Tercile of School Proportion of Students of Color

Subject	ELA						Math					
Model	Pooled T1 (1)	Dynamic T1 (2)	Pooled T2 (3)	Dynamic T2 (4)	Pooled T3 (5)	Dynamic T3 (6)	Pooled T1 (7)	Dynamic T1 (8)	Pooled T2 (9)	Dynamic T2 (10)	Pooled T3 (11)	Dynamic T3 (12)
Treat	0.110***		0.097**		-0.048*		0.127***		0.091***		-0.063**	
	(0.036)		(0.038)		(0.026)		(0.022)		(0.021)		(0.029)	
6 years before		0.004		0.133				-0.028		-0.003		
		(0.104)		(0.090)				(0.090)		(0.126)		
5 years before		-0.174		0.012		0.161***		-0.147*		-0.042		0.093*
		(0.147)		(0.041)		(0.050)		(0.088)		(0.042)		(0.048)
4 years before		-0.145		0.007		0.161***		-0.246**		0.001		0.153***
		(0.169)		(0.032)		(0.054)		(0.118)		(0.072)		(0.037)
3 years before		-0.021		-0.001		0.064		0.030		-0.019		-0.028
		(0.099)		(0.019)		(0.058)		(0.049)		(0.035)		(0.055)
2 years before		-0.050		-0.003		0.056		0.028		0.024		0.008
		(0.064)		(0.042)		(0.040)		(0.024)		(0.075)		(0.048)
1st year of fellowship		0.049		0.083***		0.025		0.026		0.064***		-0.024
		(0.065)		(0.022)		(0.036)		(0.036)		(0.021)		(0.035)
1 year after		0.011		0.133***		0.034		0.059		0.118***		0.023
		(0.060)		(0.048)		(0.026)		(0.049)		(0.025)		(0.033)
2 years after		0.017		0.143***		0.046		0.053		0.063***		-0.006
		(0.062)		(0.020)		(0.039)		(0.043)		(0.020)		(0.051)
3 years after		0.035		0.074		0.090		0.060		0.065**		-0.045
		(0.090)		(0.055)		(0.078)		(0.055)		(0.026)		(0.033)
Mean (SD) students of color	0.32(0.16)	0.32(0.16)	0.81(0.09)	0.81(0.09)	0.99(0.02)	0.99(0.02)	0.32(0.16)	0.32(0.16)	0.81(0.09)	0.81(0.09)	0.99(0.02)	0.99(0.02)
Observations	3,956	3,956	3,993	3,993	3,729	3,729	3,956	3,956	3,993	3,993	3,729	3,729
R^2	0.86958	0.87076	0.86143	0.86188	0.88240	0.88361	0.85441	0.85600	0.83909	0.83929	0.87189	0.87292
Within R^2	0.00823	0.01719	0.00656	0.00983	0.00217	0.01244	0.00761	0.01845	0.00450	0.00572	0.00289	0.01090

Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. The terciles of students of color were created by calculating the school level mean of the proportion of students of color from 2013-14 and 2014-15, where baseline race/ethnicity and poverty data was available. Each tercile includes 131 schools. All regressions include district-grade-year and school fixed effects. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels

Table 12. Difference-in-Differences Estimates of Fellowship Program on Student Achievement, by Region

	Louisiana				Kansas City				Washington D.C.			
	TWFE ELA	S&A ELA	TWFE Math	S&A Math	TWFE ELA	S&A ELA	TWFE Math	S&A Math	TWFE ELA	S&A ELA	TWFE Math	S&A Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treat	0.135***	0.142**	0.118***	0.127***	-0.029	0.006	-0.035	-0.052	-0.064**	-0.067	-0.045	-0.027
	(0.022)	(0.060)	(0.032)	(0.038)	(0.020)	(0.054)	(0.054)	(0.042)	(0.028)	(0.048)	(0.035)	(0.046)
Observations	14,214	14,214	14,214	14,214	14,160	14,160	14,160	14,160	14,224	14,224	14,224	14,224
R^2	0.88855	0.88872	0.86979	0.86996	0.88800	0.88826	0.87021	0.87045	0.88952	0.88970	0.87128	0.87153
Within R^2	0.00529	0.00680	0.00313	0.00441	0.00018	0.00246	0.00020	0.00202	0.00068	0.00229	0.00026	0.00218

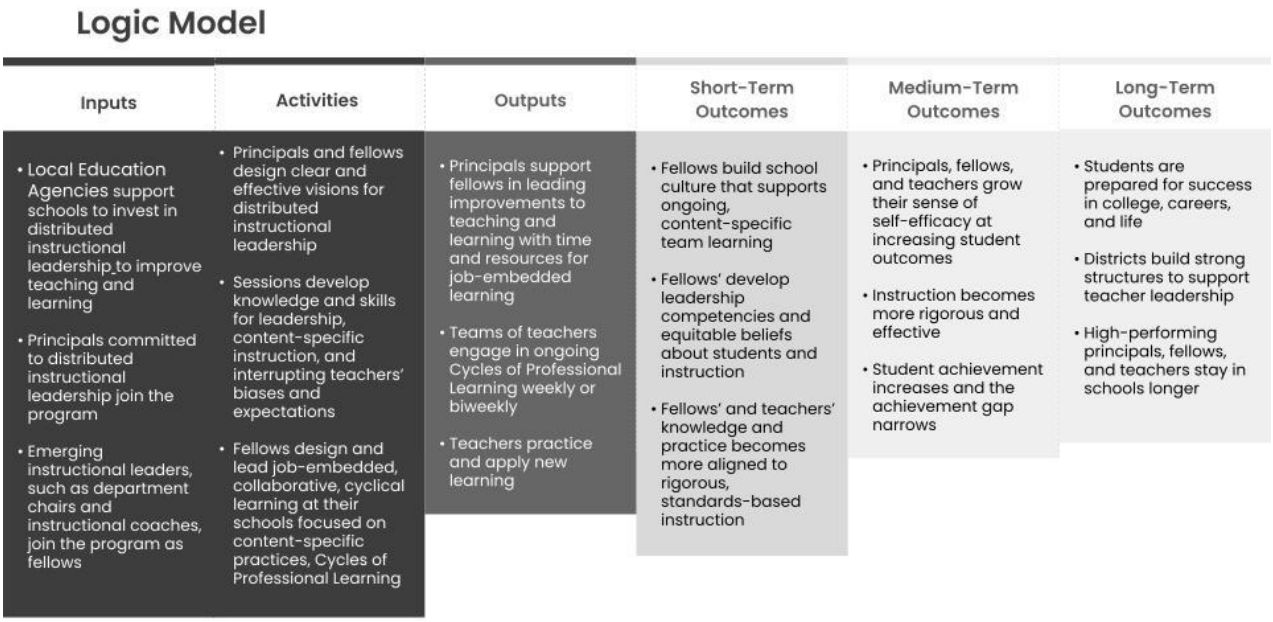
Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The S&A results are based on the procedure introduced by Sun & Abraham (2021) to account for staggered treatment timing. All regressions include district-grade-year and school fixed effects. Coefficients are statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table 13. Event Study Estimates of Fellowship Program on Student Achievement, by Region

	Louisiana				Kansas City				Washington D.C.			
	Dynamic ELA	S&A ELA	Dynamic Math	S&A Math	Dynamic ELA	S&A ELA	Dynamic Math	S&A Math	Dynamic ELA	S&A ELA	Dynamic Math	S&A Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
6 years before	0.081	0.141	0.019	0.018	0.076	0.141	-0.003	0.018	0.085	0.144	0.015	0.021
	(0.083)	(0.095)	(0.090)	(0.117)	(0.084)	(0.095)	(0.096)	(0.117)	(0.084)	(0.095)	(0.091)	(0.116)
5 years before	0.038	0.066	0.007	0.032	0.042	0.068	0.004	0.028	0.043	0.070	0.003	0.029
	(0.088)	(0.075)	(0.063)	(0.050)	(0.091)	(0.077)	(0.066)	(0.052)	(0.089)	(0.076)	(0.065)	(0.051)
4 years before	0.039	0.083	0.025	0.003	0.042	0.085	0.020	0.000	0.043	0.087	0.021	0.002
	(0.085)	(0.058)	(0.091)	(0.079)	(0.087)	(0.060)	(0.093)	(0.080)	(0.086)	(0.058)	(0.092)	(0.079)
3 years before	0.001	0.020	-0.027	-0.018	0.006	0.022	-0.035	-0.023	0.009	0.026	-0.034	-0.021
	(0.065)	(0.052)	(0.049)	(0.049)	(0.067)	(0.053)	(0.051)	(0.050)	(0.065)	(0.051)	(0.050)	(0.049)
2 years before	0.005	0.022	0.012	-0.012	0.010	0.024	0.009	-0.014	0.010	0.027	0.009	-0.012
	(0.040)	(0.035)	(0.033)	(0.042)	(0.042)	(0.036)	(0.032)	(0.041)	(0.042)	(0.036)	(0.032)	(0.041)
1st year of fellowship	0.151***	0.136**	0.093***	0.094***	-0.023	-0.015	-0.060	-0.070*	-0.057	-0.051	-0.040	0.055
	(0.047)	(0.055)	(0.033)	(0.035)	(0.058)	(0.053)	(0.053)	(0.041)	(0.058)	(0.056)	(0.053)	(0.058)
1 year after	0.158***	0.143**	0.143***	0.145***	-0.042	-0.035	0.000	-0.008	-0.025	-0.072	-0.008	-0.008
	(0.060)	(0.068)	(0.037)	(0.043)	(0.058)	(0.053)	(0.056)	(0.041)	(0.061)	(0.046)	(0.053)	(0.046)
2 years after	0.172***	0.157**	0.126***	0.127***	0.027	0.033	-0.033	-0.049	-0.039	-0.079	-0.024	-0.083
	(0.054)	(0.063)	(0.045)	(0.047)	(0.054)	(0.047)	(0.054)	(0.042)	(0.069)	(0.056)	(0.068)	(0.068)
3 years after	0.147**	0.132*	0.136***	0.137***	0.090	0.064	-0.066	-0.093	-0.043	-0.058	-0.113*	-0.117
	(0.062)	(0.070)	(0.051)	(0.052)	(0.072)	(0.089)	(0.047)	(0.067)	(0.071)	(0.080)	(0.065)	(0.071)
Observations	14,214	14,214	14,214	14,214	14,160	14,160	14,160	14,160	14,224	14,224	14,224	14,224
R^2	0.88865	0.88872	0.86985	0.86996	0.88818	0.88826	0.87027	0.87045	0.88962	0.88970	0.87137	0.87153
Within R^2	0.00618	0.00680	0.00361	0.00441	0.00173	0.00246	0.00068	0.00202	0.00157	0.00229	0.00094	0.00218

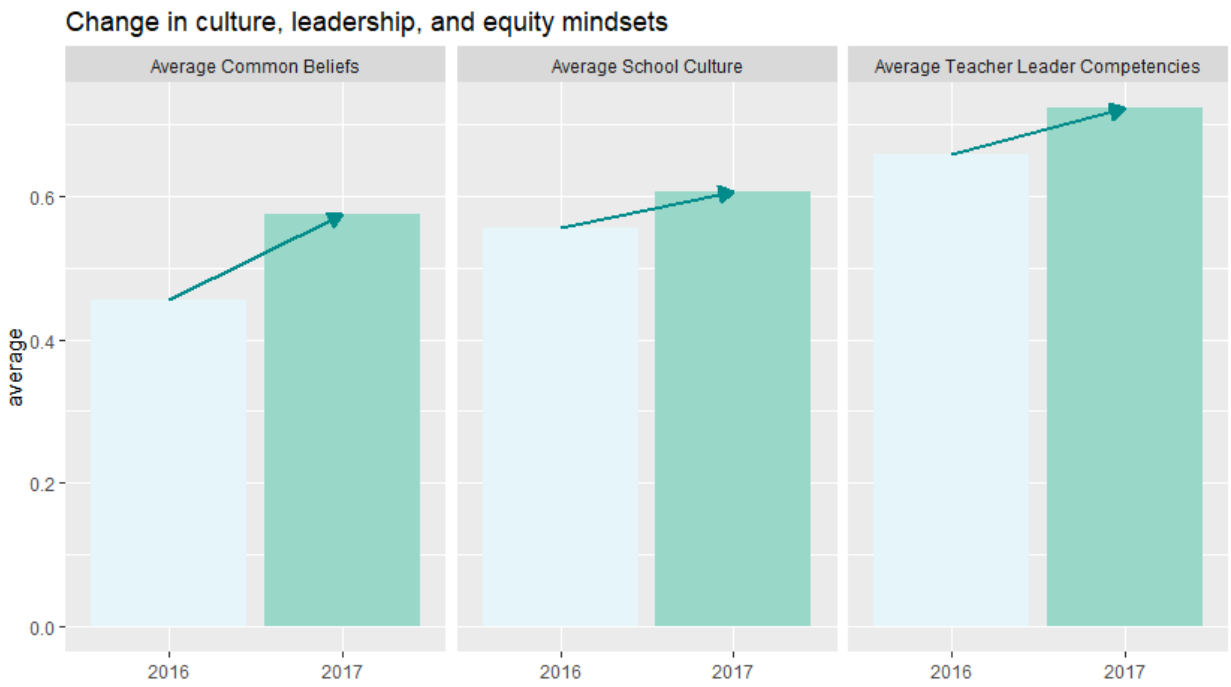
Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. *1st year of fellowship* is the 2015-16 school year for Cohort 1 and the 2016-17 school year for Cohort 2. The S&A results are based on the procedure introduced by Sun & Abraham (2021) to account for staggered treatment timing. All regressions include district-grade-year and school fixed effects. Coefficients are statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Figure 1. Leading Educators' Fellowship Program Logic Model



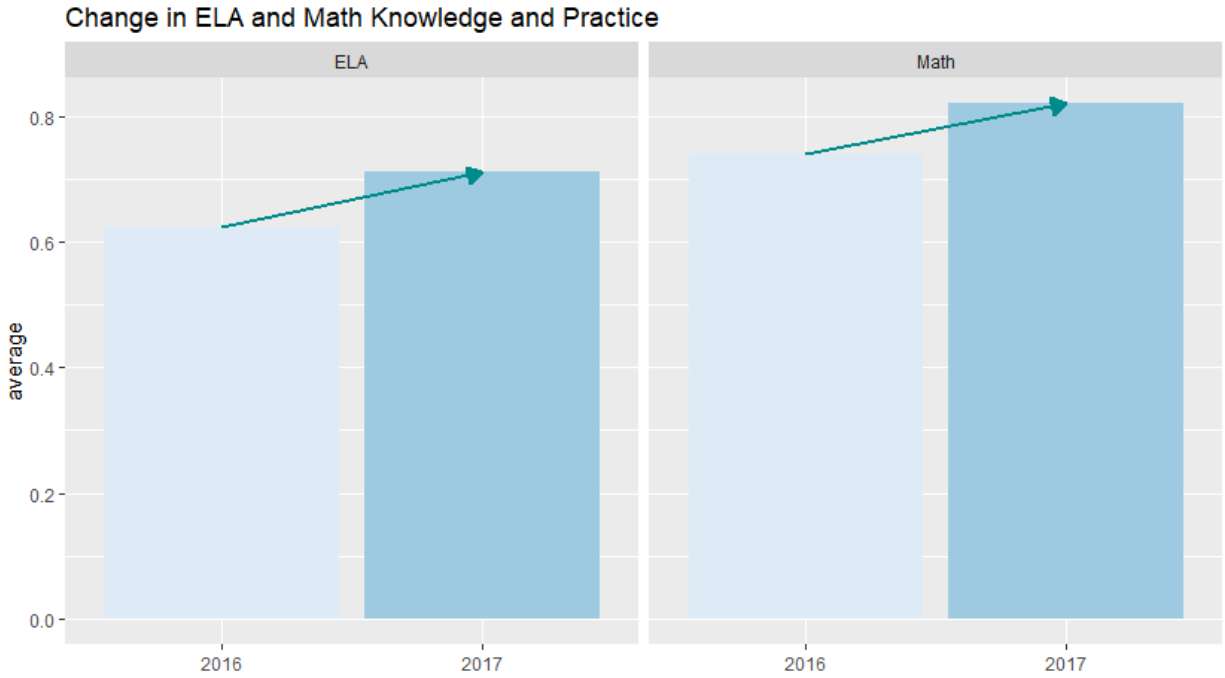
Notes: The logic model describes implementation of Leading Educators' fellowships during the 2015-16 and 2016-17 school years. Based on organizational experience and research demonstrating the value of content-specific PD, Leading Educators integrated content-specific sessions and coaching aligned to mathematics and English language arts standards for the 2016-17 school year. © Leading Educators 2017

Figure 2. Changes in Instructional Leadership Competencies, Equity Mindsets, and School Culture



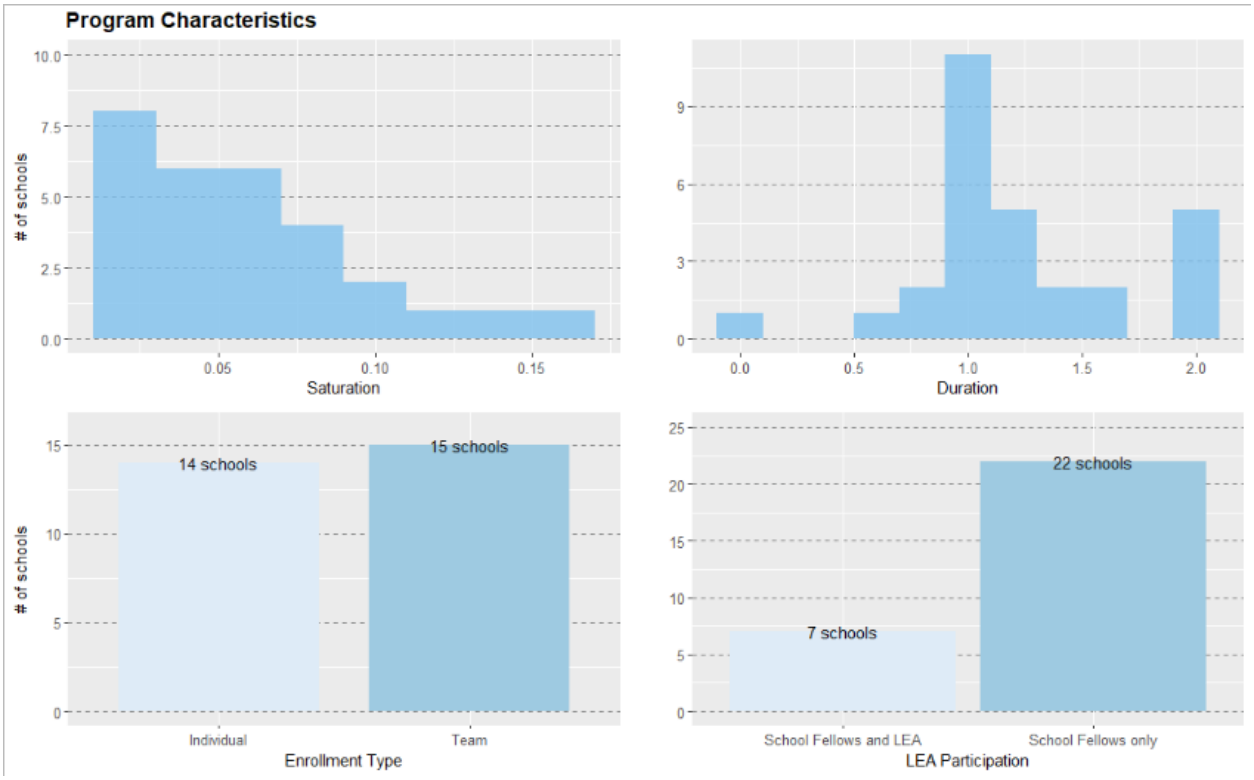
Notes: Data from Leading Educators diagnostic survey administered in the Spring of school year 2015-16 and 2016-17. The sample includes 62 fellows from Cohort 2016-17 who took the pre and posttests from a population of 112 fellows. The Average Common Beliefs is an indicator created by scaling and averaging eight prioritized equitable beliefs from the Learning for Justice's (formerly Teaching Tolerance) Common Beliefs survey (Hawley et al., n.d.). The beliefs, found in Table B1 in the appendix, were scaled to 0-1 with 1 representing more equitable beliefs and 0 representing less equitable beliefs. Teacher Leader Competencies and School Culture are indicators creating by scaling and averaging the survey items in Tables B2 and B3 correspondingly in the appendix. The survey questions were also scaled from 0-1, with 0 representing no frequency and 1 representing high frequency.

Figure 3. Fellow Knowledge and Practice, by Content Area



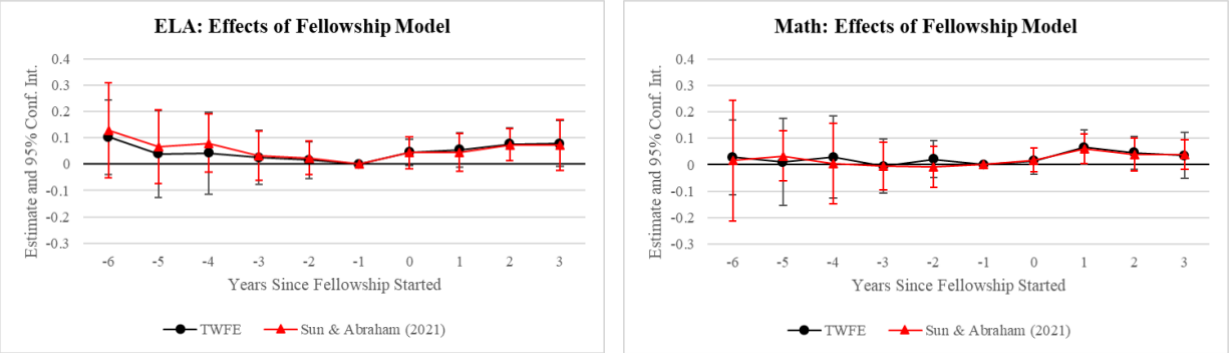
Notes: Data from Leading Educators diagnostic survey administered in the Spring of school year 2015-16 and 2016-17. The sample includes 62 fellows from Cohort 2016-17 who took the pre and posttests from a population of 112 fellows. Each bar represents a subject specific (ELA/math) survey item from the knowledge and practice assessment survey found in Table B4 in the appendix. For each knowledge item (bars 1 to 2 for ELA and 4 for math), fellows received a score between 0 and 1, with 1 being completely correct and 0 being completely not correct. For each ELA practice item (columns 3 to 5) fellows received a score from 0 to 1, with 0 representing no frequency and 1 representing high frequency. For each math practice item (columns 1 to 3), fellows received a score from 0 to 1, with 0 representing strong disagreement and 1 representing strong agreement.

Figure 4. Distribution of Program Characteristics



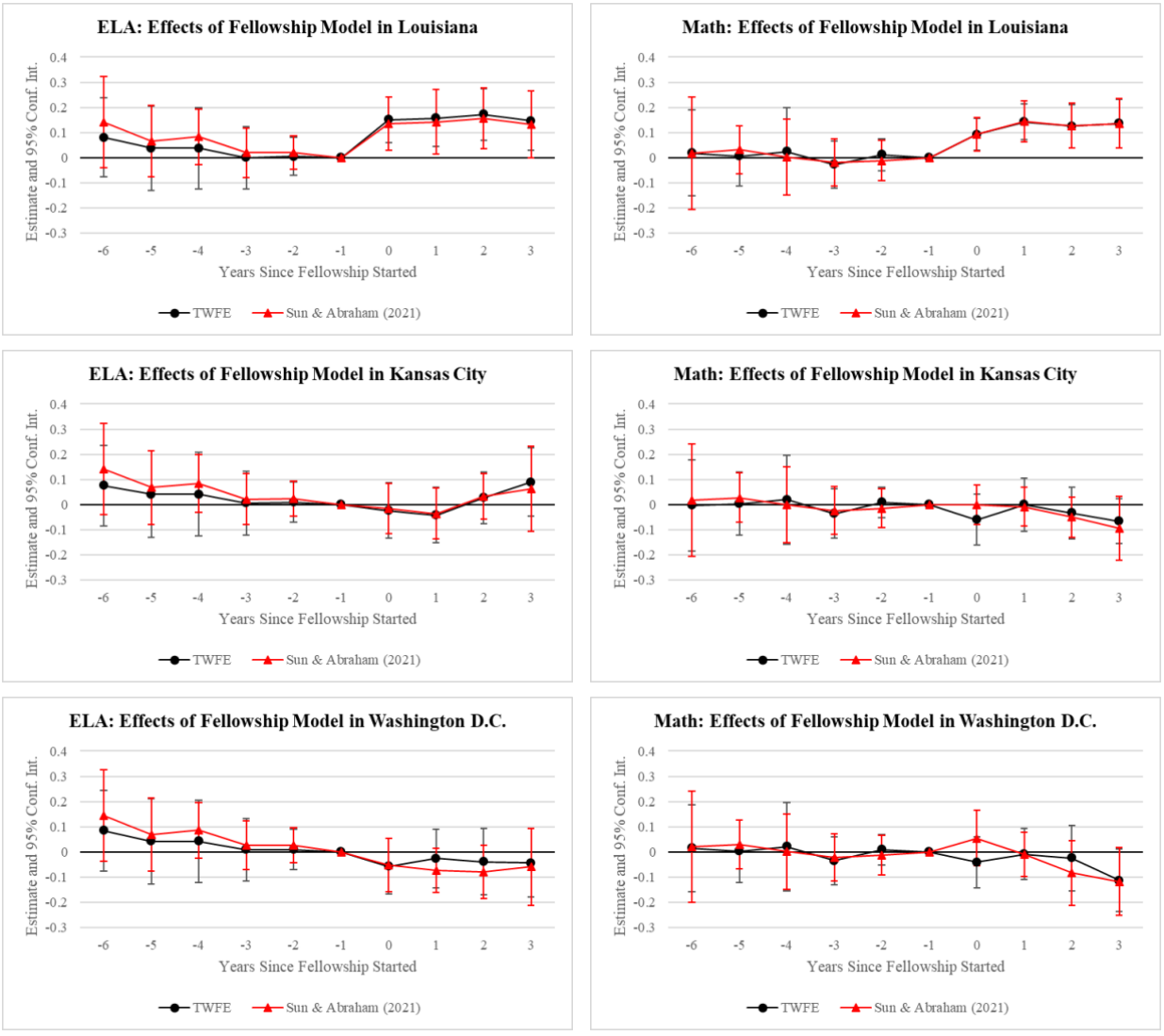
Notes: Figure shows the distributions of four characteristics that varied across schools in the fellowship model for the school-by-grade analytic sample in the program years 2015-2017, for ELA achievement outcome: Saturation (the ratio of fellows to full-time teachers in each school, expressed as a proportion); duration (years fellows remained enrolled in the program); enrollment type (whether fellows enrolled in the program as individuals or as a school team); and LEA leader participation (whether or not fellows in LEA-level roles enrolled in the program alongside fellows in school-level roles). The mean (SD) of saturation is 0.06 (0.04), and the mean (SD) of duration is 1.24 (0.48) years.

Figure 5. Event Study Estimates of Fellowship Program Effects on Student Achievement



Notes: Figure shows the average effect of the fellowship program (with 95% confidence intervals) on the proportion of students scoring proficient or advanced in ELA and Math in the pre- and post- treatment periods (relative to the year immediately before the first year of treatment). See Table 4 for a complete summary of estimates upon which these event study estimates are based. Year 0 is the first year of the fellowship (2015-16 for Cohort 1; 2016-17 for Cohort 2).

Figure 6. Event Study Estimates of Fellowship Program Effects on Student Achievement, by Region



Notes: Figure shows the average effect of the fellowship program (with 95% confidence intervals) on the proportion of students scoring proficient or advanced in ELA and Math in the pre- and post- treatment periods (relative to the year immediately before the first year of treatment). See Table 14 for a complete summary of estimates upon which these event study estimates are based. Year 0 is the first year of the fellowship (2015-16 for Cohort 1; 2016-17 for Cohort 2).

Appendix A: Tables & Figures

Table A1. Baseline Characteristics, by Treatment Status and Region

	Kansas City			Louisiana			Washington D.C.		
	Comparison	Treatment	Pooled	Comparison	Treatment	Pooled	Comparison	Treatment	Pooled
ELA proficiency	0.52 (0.21)	0.42 (0.18)	0.52 (0.21)	0.64 (0.2)	0.77 (0.21)	0.65 (0.2)	0.24 (0.17)	0.25 (0.19)	0.24 (0.17)
Math proficiency	0.37 (0.21)	0.31 (0.19)	0.37 (0.21)	0.56 (0.22)	0.68 (0.17)	0.57 (0.22)	0.23 (0.17)	0.22 (0.16)	0.23 (0.17)
Hispanic	0.14 (0.18)	0.19 (0.27)	0.14 (0.18)	0.09 (0.12)	0.2 (0.08)	0.09 (0.12)	0.12 (0.19)	0.11 (0.15)	0.11 (0.19)
Black	0.34 (0.32)	0.57 (0.32)	0.35 (0.32)	0.72 (0.3)	0.37 (0.19)	0.71 (0.3)	0.81 (0.25)	0.81 (0.24)	0.81 (0.25)
White	0.46 (0.33)	0.2 (0.24)	0.45 (0.33)	0.15 (0.22)	0.37 (0.2)	0.16 (0.22)	0.04 (0.12)	0.06 (0.11)	0.05 (0.11)
Asian	0.02 (0.03)	0.01 (0.02)	0.02 (0.03)	0.02 (0.04)	0.02 (0.02)	0.02 (0.04)	0.01 (0.02)	0 (0.01)	0.01 (0.02)
Poverty	0.25 (0.11)	0.31 (0.07)	0.25 (0.11)	0.26 (0.07)	0.24 (0.07)	0.26 (0.07)	0.3 (0.08)	0.25 (0.08)	0.29 (0.08)
Modified Poverty	0.25 (0.13)	0.33 (0.1)	0.25 (0.13)	0.33 (0.1)	0.39 (0.12)	0.33 (0.1)	0.32 (0.08)	0.27 (0.08)	0.31 (0.08)
English Learners	0.1 (0.16)	0.13 (0.25)	0.1 (0.17)	0.06 (0.08)	0.09 (0.08)	0.06 (0.08)	0.07 (0.11)	0.06 (0.08)	0.07 (0.1)
Special Education	0.11 (0.05)	0.1 (0.06)	0.11 (0.06)	0.1 (0.04)	0.06 (0.01)	0.1 (0.04)	0.12 (0.05)	0.14 (0.06)	0.13 (0.05)
Student-Teacher Ratio	14.83	13.22	0.04	15.39	14.89	0.1	14.15	11.71	0
Magnet	0.04	0.15	14.75	0.1	0	15.38	0	0	13.73
Charter	0.13	0.46	0.14	0.44	0.82	0.45	1	1	1
Grade 3	0.24	0.23	0.24	0.21	0.29	0.21	0.18	0.23	0.19
Grade 4	0.24	0.23	0.24	0.2	0.24	0.2	0.17	0.18	0.17
Grade 5	0.22	0.19	0.22	0.2	0.18	0.2	0.18	0.18	0.18
Grade 6	0.15	0.19	0.15	0.13	0.12	0.13	0.17	0.15	0.16
Grade 7	0.08	0.12	0.08	0.14	0.12	0.13	0.16	0.13	0.15
Grade 8	0.08	0.04	0.07	0.12	0.06	0.12	0.15	0.13	0.15
Schools	159	8	167	214	5	219	55	11	66
School-Grade Observations	514	26	540	798	17	815	186	39	225

Notes: Each cell reports mean (standard deviation), except for grade-level, magnet, and *c*, which reports proportions. Data are based on school-grade-year analytic sample for ELA and Math achievement outcome during the baseline (i.e., pre-treatment) year 2014-15, except for English Learners and Special Education where baseline data was only available in the 2013-14 school year. Baseline data were unavailable for 5 treatment schools; therefore, their baseline characteristics do not appear in the table above. *ELA* and *Math Proficiency* is based on data reported by EDFacts (via the Urban Institute's Education Data Portal) and is measured as the midpoint of the range used to report the share of students achieving "proficient" or "advanced" levels as defined by each State Education Agency on the state's English language arts (ELA) and math assessments. While it may appear baseline proficiency in Louisiana was significantly higher than other regions, this

difference is more reflective of differing proficiency bars; mean 4th grade scores on the National Assessment for Education Progress in DC and Louisiana were identical in 2015, and results in Missouri were slightly higher but around the national average.

Table A2. Effects of Fellowship Program, by Saturation

	Pooled ELA (1)	Dynamic ELA (2)	Pooled Math (3)	Dynamic Math (4)
Treat	0.202***		0.188***	
	(0.039)		(0.039)	
Saturation*Treat	-3.021***		-2.856***	
	(0.682)		(0.858)	
6 years before		0.067		-0.003
		(0.081)		(0.098)
5 years before		0.035		0.005
		(0.088)		(0.065)
4 years before		0.035		0.021
		(0.085)		(0.091)
3 years before		0.001		-0.031
		(0.065)		(0.051)
2 years before		0.006		0.013
		(0.039)		(0.032)
1st year of fellowship		0.217***		0.166***
		(0.052)		(0.048)
1 year after		0.239***		0.208***
		(0.081)		(0.066)
2 years after		0.232***		0.202***
		(0.072)		(0.051)
3 years after		0.175***		0.224***
		(0.050)		(0.046)
Saturation*1st year of fellowship		-3.206***		-2.795***
		(0.875)		(0.993)
Saturation*1 year after		-3.487***		-2.644***
		(0.854)		(0.967)
Saturation*2 years after		-2.914***		-2.928***
		(0.751)		(0.958)
Saturation*3 years after		-1.798		-3.793***
		(1.353)		(1.029)
Observations	14,454	14,454	14,454	14,454
R^2	0.88994	0.89007	0.87173	0.87183
Within R^2	0.00473	0.00588	0.00319	0.00396

Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the average saturation of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. Saturation represents the effect of increasing the average saturation of fellows for each full-time teacher by one percentage point. *Saturation* is measured as the ratio of fellows in school-level roles to that school's overall number of full-time teaching staff. The mean (standard deviation) of saturation is 0.06 (0.04). All regressions include district-grade-year and school fixed effects. Coefficients are statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table A3. Effects of Fellowship Program, by Enrollment Type

	Pooled ELA (1)	Dynamic ELA (2)	Pooled Math (3)	Dynamic Math (4)
Treat	0.030		0.057	
	(0.067)		(0.068)	
Team*Treat	0.006		-0.053	
	(0.077)		(0.076)	
6 years before		0.103		0.024
		(0.077)		(0.094)
5 years before		0.040		0.007
		(0.089)		(0.068)
4 years before		0.043		0.025
		(0.085)		(0.092)
3 years before		0.026		-0.011
		(0.055)		(0.041)
2 years before		0.015		0.020
		(0.039)		(0.030)
1st year of fellowship		0.047		0.030
		(0.046)		(0.045)
1 year after		0.046		0.077*
		(0.038)		(0.042)
2 years after		0.069*		0.092**
		(0.041)		(0.040)
3 years after		0.070		0.063
		(0.050)		(0.074)
Team*1st year of fellowship		-0.004		-0.031
		(0.085)		(0.086)
Team*1 year after		0.015		-0.022
		(0.092)		(0.078)
Team*2 years after		0.015		-0.098
		(0.078)		(0.070)
Team*3 years after		0.016		-0.060
		(0.085)		(0.106)
Observations	14,454	14,454	14,454	14,454
R^2	0.88950	0.88964	0.87143	0.87153
Within R^2	0.00076	0.00200	0.00085	0.00168

Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program for the sample of schools where fellows applied as individuals on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. Team enrollment is the effect of going from individual to team enrollment, where $Team = 1$ if fellows in school *s* enrolled as a team and 0 if fellows in school *s* enrolled as individuals. All regressions include district-grade-year and school fixed effects. Coefficients are statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table A4. Effects of Fellowship Program, by Duration

	Pooled ELA (1)	Dynamic ELA (2)	Pooled Math (3)	Dynamic Math (4)
Treat	-0.092**		-0.062	
	(0.038)		(0.041)	
Time Enrolled*Treat	0.101***		0.073**	
	(0.029)		(0.035)	
6 years before		0.096		0.024
		(0.080)		(0.098)
5 years before		0.037		0.008
		(0.089)		(0.065)
4 years before		0.039		0.026
		(0.084)		(0.090)
3 years before		0.014		-0.015
		(0.064)		(0.049)
2 years before		0.011		0.019
		(0.040)		(0.033)
1st year of fellowship		-0.085		-0.075
		(0.070)		(0.061)
1 year after		-0.070		-0.033
		(0.085)		(0.051)
2 years after		-0.052		-0.009
		(0.074)		(0.073)
3 years after		-0.052		-0.099
		(0.074)		(0.072)
Time Enrolled*1st year of fellowship		0.102***		0.071*
		(0.030)		(0.037)
Time Enrolled*1 year after		0.098**		0.078***
		(0.045)		(0.028)
Time Enrolled*2 years after		0.101***		0.044
		(0.034)		(0.049)
Time Enrolled*3 years after		0.101***		0.103**
		(0.028)		(0.046)
Observations	14,454	14,454	14,454	14,454
R^2	0.88990	0.89003	0.87156	0.87166
Within R^2	0.00438	0.00552	0.00190	0.00269

Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the average duration of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. Time enrolled represents the effect of increasing the average duration by 0.1 years (1.2 months). The mean (standard deviation) of time enrolled is 1.22 (0.46) years. All regressions include district-grade-year and school fixed effects. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table A5. Effects of Fellowship Program, by LEA Leader Participation

	Pooled ELA (1)	Dynamic ELA (2)	Pooled Math (3)	Dynamic Math (4)
Treat	-0.051*		-0.017	
	(0.030)		(0.042)	
TeachLEA*Treat	0.120**		0.066	
	(0.053)		(0.069)	
6 years before		0.108		0.031
		(0.077)		(0.096)
5 years before		0.045		0.013
		(0.087)		(0.063)
4 years before		0.047		0.031
		(0.082)		(0.088)
3 years before		0.023		-0.007
		(0.060)		(0.045)
2 years before		0.016		0.022
		(0.039)		(0.032)
1st year of fellowship		-0.034		-0.038
		(0.057)		(0.051)
1 year after		-0.039		0.023
		(0.059)		(0.053)
2 years after		-0.014		0.015
		(0.059)		(0.059)
3 years after		0.003		-0.027
		(0.067)		(0.082)
TeachLEA*1st year of fellowship		0.117*		0.079
		(0.060)		(0.071)
TeachLEA*1 year after		0.136**		0.062
		(0.067)		(0.069)
TeachLEA*2 years after		0.133**		0.043
		(0.057)		(0.074)
TeachLEA*3 years after		0.107*		0.092
		(0.062)		(0.098)
Observations	14,454	14,454	14,454	14,454
R^2	0.88972	0.88987	0.87143	0.87152
Within R^2	0.00274	0.00411	0.00090	0.00156

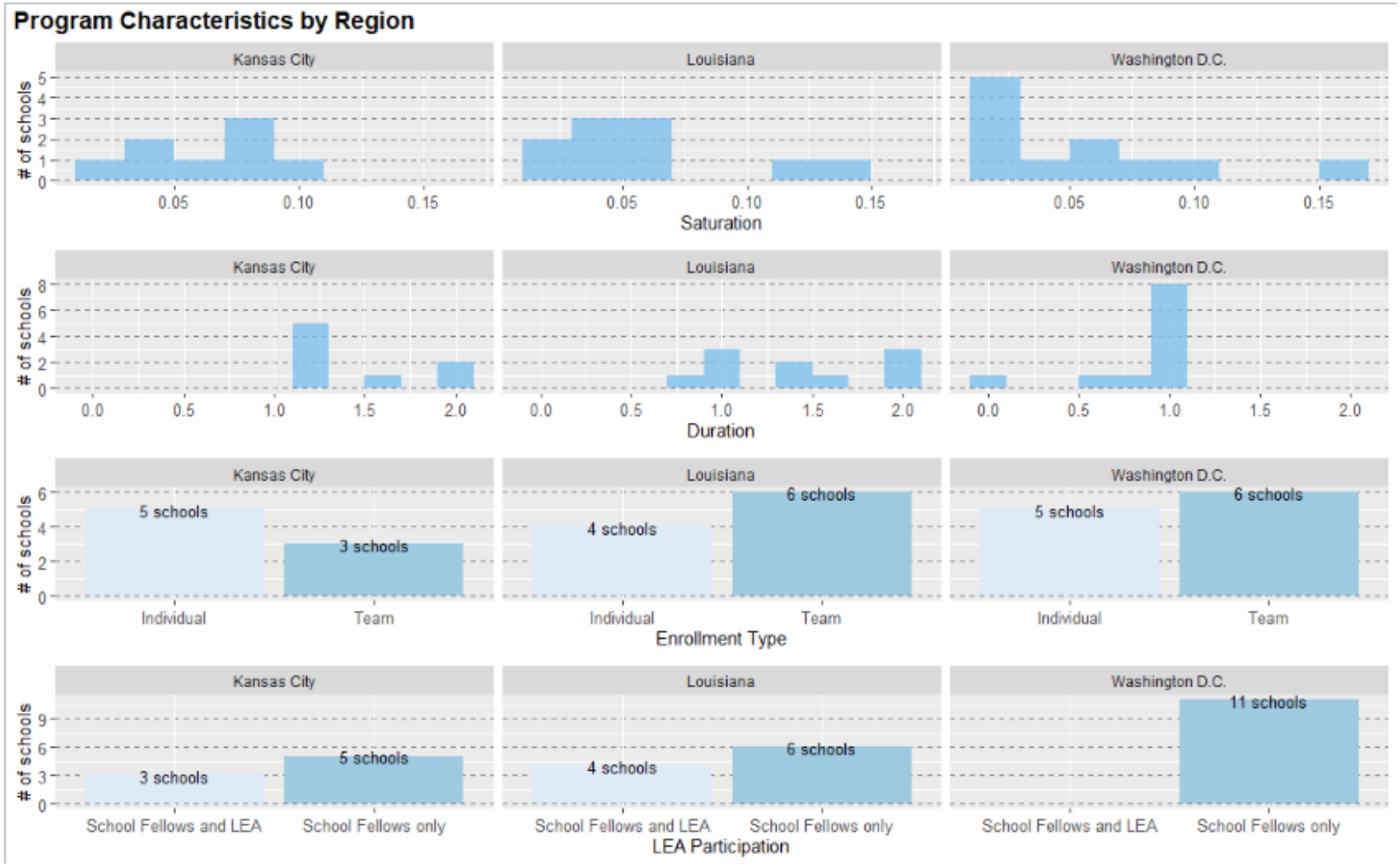
Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program for the sample of schools where only fellows with school-level roles participated on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. $TeachLEA = 1$ if both school-level fellows and LEA-level fellows were enrolled in the program in school *s* and 0 if only school-level fellows were enrolled in the program in school *s*. All regressions include district-grade-year and school fixed effects. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Table A6. Effects of Fellowship Program on ELA and Math, Main and Heterogeneity Samples

Subject	ELA				Math			
Model	Pooled main	Pooled hetero	Dynamic main	Dynamic hetero	Pooled main	Pooled hetero	Dynamic main	Dynamic hetero
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat	0.034	0.034			0.029	0.030		
	(0.042)	(0.040)			(0.044)	(0.042)		
6 years before			0.102	0.103			0.028	0.028
			(0.075)	(0.072)			(0.097)	(0.092)
5 years before			0.039	0.040			0.010	0.009
			(0.087)	(0.084)			(0.064)	(0.061)
4 years before			0.042	0.043			0.029	0.029
			(0.082)	(0.079)			(0.088)	(0.084)
3 years before			0.025	0.025			-0.006	-0.006
			(0.054)	(0.052)			(0.041)	(0.039)
2 years before			0.015	0.015			0.021	0.022
			(0.037)	(0.036)			(0.031)	(0.029)
1st year of fellowship			0.045*	0.046*			0.015	0.015
			(0.026)	(0.025)			(0.021)	(0.020)
1 year after			0.054	0.055*			0.066**	0.064**
			(0.034)	(0.032)			(0.026)	(0.025)
2 years after			0.076**	0.076**			0.046*	0.049**
			(0.033)	(0.031)			(0.023)	(0.022)
3 years after			0.078*	0.077*			0.035	0.038
			(0.046)	(0.045)			(0.032)	(0.029)
Observations	14,454	11,678	14,454	11,678	14,454	11,678	14,454	11,678
R^2	0.88950	0.88621	0.88964	0.88637	0.87137	0.86628	0.87144	0.86636
Within R^2	0.00076	0.00088	0.00195	0.00226	0.00043	0.00052	0.00098	0.00114

Notes: Each column represents a separate regression. Coefficients with robust standard errors (clustered at the school level) are reported. Data are from 2009-10 through 2018-19 school years. *Treat* is the pooled difference-in-differences estimate and represents the estimated impact of the fellowship program on the proportion of students who are proficient or advanced in the post-treatment period (2015-16 to 2018-19 school years). The year-specific effects represent the event study estimates and are relative to the baseline pre-treatment year. The main sample is larger than the heterogeneity sample because the heterogeneity sample includes schools with available poverty (modified poverty) and race/ethnicity data for the baseline years used to construct the poverty (modified poverty) and students of color terciles in the heterogeneity analysis (see tables 9-11). All regressions include district-grade-year and school fixed effects. Coefficients statistically significant at the 10% (*), 5% (**), and 1% (***) levels.

Figure A1. Distribution of Program Characteristics, by Region



Notes: Figure shows the distributions of four characteristics that varied across schools in the fellowship model for the school-by-grade analytic sample in the program years 2015-2017, for ELA achievement outcome: Saturation (the ratio of fellows to full-time teachers in each school, expressed as a proportion); duration (years fellows remained enrolled in the program); enrollment type (whether fellows enrolled in the program as individuals or as a school team); and LEA leader participation (whether or not fellows in LEA-level roles enrolled in the program alongside fellows in school-level roles). The mean (SD) of saturation is 0.06 (0.04), and the mean (SD) of duration is 1.24 (0.48) years.

Appendix B: Surveys and Rubrics

Table B1. Common Beliefs Survey

(Adapted, with permission, from a self-reflection tool created by Learning for Justice (formerly Teaching Tolerance; Hawley et al., n.d.))

Beliefs about Equity: Please rate your agreement with the following statements.

	Strongly Agree	**Agree**	**Somewhat Agree**	**Neutral**	**Somewhat Disagree**	**Disagree**	**Strongly Disagree**
I don't think of my students in terms of their race or ethnicity. I am color blind when it comes to my teaching.							
The gap in achievement among students of different races is about poverty not race.							
Teachers should adapt their teaching to the distinctive cultures of African American, Latino, Asian, and Native American students.							
In some cultures, students are embarrassed to speak in front of others, and so I take this into account and don't call on those students in class.							
When students come from homes where educational achievement is not a high priority, they often don't do their homework and their parents don't come to school events. This lack of parental support undermines my efforts to teach these students.							
It is not fair to ask students who are struggling with English to take on challenging academic assignments.							
I believe I should reward students who try hard, even if they are not doing well in school. Building their self-esteem is important.							
I try to keep in mind the limits of my students' ability and give them assignments I know they can do so they do not become discouraged.							
Students of different races and abilities often have different learning styles, and good teachers match their instruction to those styles.							

Grouping students of different levels of achievement for instruction may benefit some students, but it can undermine the progress that could otherwise be made by higher achieving students.							
Before students are asked to engage in complex learning tasks, they need to have a solid grasp of basic skills.							
With all the pressures to raise student achievement, finding and using examples for the cultural, historic and everyday lived experiences of my students takes away (or could take away) valuable time from teaching and learning what matters most.							
Talking about race with my colleagues could open up a can of worms; little good is likely to come from it.							

Table B2. Teacher leader competencies survey
(Developed by Leading Educators)

Teacher leader competencies: How often do you carry out the following actions?

Statement	**Almost Never**	**Sometimes**	**Frequently**	**Almost Always**	**Always**	**Not Applicable/ Do Not Know**
Models the belief that, regardless of circumstances, all children can master rigorous material						
Acknowledges and confronts racial and other biases in self and others						
Accurately senses and seeks to understand colleagues' preferences, emotions, and perspectives						
Accepts criticism and feedback and takes responsibility for actions						

Adjusts behaviors to respect colleagues' preferences, emotions, and perspectives						
Creates a sense of urgency by ensuring team members see the need for change and the importance of immediate action						
Makes decisions to drive high-quality results while respecting the values and capacity of teammates and school (potential subtract)						
Invites diverse stakeholders (considering students, community members, school leaders, and colleagues) to provide input on and participate in project planning and implementation						
Shifts cognitive lift to teammates by asking open-ended non-rhetorical questions and including think time						
When giving feedback, makes specific, non-judgmental, and factual statements						
Encourages unfiltered discussion to explore differing opinions and reach shared decisions						
Analyzes evidence of teaching and learning with team, celebrating success and making adjustments as needed						
Develops strong relationships and actively renews and repairs relationships as necessary.						

Table B3. School culture survey
(Developed by Leading Educators)

School culture How often do the following actions occur?

Statement	Almost Never	Sometimes	Frequently	Almost Always	Always	Not Applicable/ Do Not Know
Teachers are focused on creating equity for all students through rigorous content.						
Teachers tailor their instructional practices to meet their students' individual learning needs.						
Teachers share norms and values.						
Teachers are honest about their growth areas and ask for help when needed.						
Teachers continually deepen their content knowledge and pedagogy.						
Teachers practice and get feedback on new instructional and planning skills.						
Professional learning opportunities include opportunities to experience, reflect, build shared language, and apply new knowledge.						
Meetings and professional learning opportunities target specific goals to improve teacher and student learning.						
Teachers trust and respect one another.						
Teachers talk to each other and gather ideas about the specific challenges they face in their own classrooms.						
Teachers stick to evidence and data when making statements and decisions.						
Teachers share and give feedback on each other's lesson plans.						

Teachers share and give feedback on each other's teaching practice.						
Teachers consider evidence of student learning when planning for instruction.						
Teachers rely on evidence when deciding whether to adopt new instructional materials or practices.						
Teachers provide students with many opportunities to participate in classroom discussions.						
Teachers consider student input or preferences when planning instructional units.						
Teachers seek out strategies for making classroom content engaging for all students.						

Table B4. Knowledge and practice assessment survey
(Adapted from an assessment created by Student Achievement Partners in 2014)

MATH KNOWLEDGE AND PRACTICE [ECE AND K-2]

This section is focused on standards-based instructional planning and practice. The purpose of this section is to get a sense of of what teachers know and are able to do before planning professional learning for the year. These questions are designed to get accurate data of strengths, preconceived notions, misunderstandings, misconceptions, and knowledge gaps. If you have resources you regularly use in your instructional planning, feel free to consult them.

This section is intentionally challenging. Do not be discouraged if you do not know all or most of the answers. If "I don't know" is the most accurate answer, please use it. Your answers here will not be used to evaluate or judge you, but instead to get an accurate gauge of planning and practice of standards-based instruction in your school. This information will ensure professional learning cycles are aligned to your practice.
Which of the following belongs to the major work of the indicated grade?

Grade	**Select all that apply**

K	Compare numbers	Tell and write time from analog and digital clocks to the nearest five minutes using a.m. and p.m.	Understand meaning of addition and subtraction	Develop understanding of fractions as numbers	I don't know
1	Add and subtract within 20	Measure lengths indirectly and by iterating length units	Extend understanding of fraction equivalence and ordering	Identify arithmetic patterns (including patterns in the addition or multiplication tables) and explain them using properties of operations	I don't know
2	Identify line of symmetry in two dimensional figures	Understand place value	Apply and extend previous understandings of multiplication and division to multiply and divide fractions	Represent and solve problems involving addition	I don't know

Please rate your agreement with the following statements.

Statement	**Strongly Agree**	**Agree**	**Somewhat Agree**	**Neutral**	**Somewhat Disagree**	**Disagree**	**Strongly Disagree**
I regularly apply rigor in standards and assessment in my instruction.							
I regularly apply deep knowledge of mathematical content and pedagogy in my grade band in my instruction.							
I regularly apply the standards for mathematical practice in my instruction.							

ELA KNOWLEDGE AND PRACTICE [ECE AND LOWER ELEMENTARY K-2 ONLY]

This section is focused on standards-based instructional planning and practice. The purpose of this section is to get a sense of of what teachers know and are able to do before planning professional learning for the year. These questions are designed to get accurate data of strengths, preconceived notions, misunderstandings, misconceptions, and knowledge gaps. If you have resources you regularly use in your instructional planning, feel free to consult them.

This section is intentionally challenging. Do not be discouraged if you do not know all or most of the answers. If "I don't know" is the most accurate answer, please use it. Your answers here will not be used to evaluate or judge you, but instead to get an accurate gauge of planning and practice of standards-based instruction in your school. This information will ensure professional learning cycles are aligned to your practice.

In a typical lesson, please respond about the percentage of your time you are engaging students in the following activities during class. Percentages do not need to add up to 100, as some items may overlap.

	Never use	**1-25% of a typical lesson**	**26-50% of a typical lesson**	**51-75% of a typical lesson**	**76-100% of a typical lesson**
Use of a single grade-level text for whole-class reading, writing, and/or discussion					
Use of leveled readers to support struggling students in place of the grade-level text other students are reading in class					
In-class writing assignments in response to or about texts					

Which of the following approaches for selecting texts for reading aligns with your state's English language arts and literacy standards? Check all that apply.

1. Using abridged or adapted versions of complex texts for struggling readers
2. Assigning complex texts that all students in a class are required to read
3. Selecting texts for individual students based on their reading level
4. Selecting texts for a class based on qualitative factors like knowledge demands, as well as quantitative factors like word and sentence length
5. Other approach (please describe):
6. I don't know

Mr. Jones is developing a lesson plan to go with the text, "Lost Penguin Back in his Natural Habitat." How could Mr. Jones provide the appropriate scaffolds so that all students - including those who read below grade-level - have opportunities to engage in the work of the lesson in a way that best aligns with your state's English language arts and literacy standards?

1. He could rewrite the text and substitute more complex text and difficult vocabulary with easier words and phrases.
2. He could create a podcast or audio recording of the passage for students to listen to as they read along.
3. He could build background knowledge by providing a summary of the text.
4. He could group students homogeneously and give the English Language Learners a simpler text on the same subject.
5. I don't know

Table B5. Leading Educators Selection Process

Application Process Overview

In order to apply, fellows must have at least two years experience and have received a rating of Effective or Highly Effective on last year's Evaluation. All candidates complete the same process, and each phase is scored by assessors using rubrics who pass a norming and training exercise.

1. **Written Application:** applicants submit a short application answering two questions:
    0. The Leading Educators Fellowship is a rigorous and challenging two-year program that empowers teacher leaders to expand their impact. What do you hope to learn and develop through your participation in the Fellowship?
    1. Please describe your teacher leader responsibilities for next year. How do you hope to impact the teachers and students at your school in this role?
    2. Describe a time in your past two years as an educator when you set a goal with your students and struggled to meet it. How did you respond, and what do you think contributed to your success or failure?
2. **Instructional Assessment:** applicants submit a student data essay describing results with students and are observed teaching a lesson in the content area and submit an evidence-based reflection on what students learned and how it prepared them for mastery of grade-level content.
3. **Interview and Role Play:** applicants participate in an interview where they answer questions about their content area and participate in a short role play. Sample questions include:
    0. What are the key priorities of the school in this content area? How would you address these priorities as a teacher leader?
    1. In the last year, how have you deliberately grown in your instructional practice? How do you know?
    2. The applicant participates in a short role play giving feedback to a teacher to assess their content expertise and how they manage the relationship.

	Written Application	Instructional Assessment	Interview and Role Play
Equity: Models the belief that all children can master rigorous material			
Growth: Actively seeks opportunities to leverage strengths and develop growth areas			
Opportunities to Lead: Clear charge from the principal in terms of a leadership role or responsibility for the next two years. Responsible for leading at least two adults.			
Results: Works diligently and purposefully to reach results without lowering expectations			

Vision: Clearly communicates vision of success for students and teachers			
Assess: Analyzes and reflects on student achievement data against benchmarks towards end of year goals			
Instructional Expertise: Achieves strong results with students			
Relationship Management Appropriately matches leadership styles to individual and contextual needs by identifying the skill level and motivation of colleagues			
Community Collaborates with colleagues to increase the collective impact on student success Supports, celebrates, and challenges colleagues			
Self-Management Identifies emotional triggers and manages reactions to conflict and stressful situations			
Plan: Analyzes context to identify the highest-need annual and interim priorities with clear links to vision of success for students and teachers			