Research Report

KATA MIHALY, ISAAC M. OPPER, LUCAS GREER

The Impact and Implementation of the Chicago Collaborative Teacher Professional Development Program

Teachers, like the students they serve, never stop learning. In-service teacher professional development (PD) gives educators opportunities over the course of the year to learn more about pedagogy and improve their own instruction methods to boost students' academic and social and emotional outcomes. On average, teachers participate in ten days of in-service PD during the school year. Opportunities might include taking a formal course, attending relevant conferences, engaging in teacher learning networks or peer mentoring, or taking part in observational visits to other schools and organizations (Sellen, 2016).

Given the amount of time spent in these activities, teacher PD comes at a high cost to districts. In fact, according to some estimates, districts can pay as much as $18,000 per teacher per year (Hill, 2015; The New Teacher Project, 2015). Despite this significant financial investment, research on the impact of teacher PD on teacher instruc-

KEY FINDINGS

- The Leading Educators Chicago Collaborative program, a teacher professional development program aligned to Common Core State Standards, was successfully delivered despite the challenges posed by the coronavirus pandemic.
- Implementation varied across districts and schools, and Chicago Public Schools used a format closest to the original Leading Educators model.
- Students attending schools that were randomly assigned to the Chicago Collaborative program had statistically significantly higher test scores compared with students attending schools that were randomly assigned to the control group.
- Improvements in test score varied by subgroups, with larger effects in middle school. However, these differences were not statistically significant.
- The Leading Educators Chicago Collaborative was effective at improving test scores and is a promising model to study on a larger scale.

tional practices and student education outcomes has been mixed. In addition, there are only a few studies that examine the impact of teacher PD using rigorous empirical evaluation designs (such as randomized control trials) and consider PD across multiple contexts (public versus charter schools).

In this research, we evaluate the Chicago Collaborative, a teacher PD program that is aligned to Common Core State Standards (CCSS) and implemented by Leading Educators. Leading Educators is a national nonprofit organization that partners with districts and charter management organizations to help teachers develop the leadership skills that they need to successfully transition from leading students to leading their peers. The Chicago Collaborative is based on Leading Educators' theory of change—specifically, that developing teacher-leaders to lead job-embedded content-specific cycles of professional learning aligned to college- and career-ready standards will result in successful outcomes, including (1) high-performing teachers retained in their jobs, (2) an increase in teacher pedagogical content knowledge (how to teach), and (3) a closing of the opportunity gap in high-needs schools. An earlier version of the training model used by Leading Educators has prior evidence of effectiveness (Mihaly, Master, and Yoon, 2015) and includes design features that are supported by theoretical evidence (Merchie et al., 2018).

Abbreviations

CCSS	Common Core State Standards
COVID-19	coronavirus disease 2019
CPS	Chicago Public Schools
ELA	English language arts
ELL	English language learner
FRPL	free or reduced-price lunch
IEP	individualized education plan
MAP	Measures of Academic Progress
NWEA	Northwest Evaluation Association
PARCC	Partnership for Assessment of Readiness for College and Careers
PD	professional development
SD	standard deviation
TOT	treatment-on-the-treated

How the Research Was Conducted

Our research took place during the 2018–2019 and 2019–2020 school years in 40 schools across three school districts in the Chicago area. We sought answers to two interrelated questions:

- How is the Chicago Collaborative PD program implemented?
- Did the Chicago Collaborative PD program impact student achievement?

Figure 1 summarizes the activities we undertook to answer these questions. Details about our data collection and analysis methods are presented in the main body of the report and the accompanying appendix.

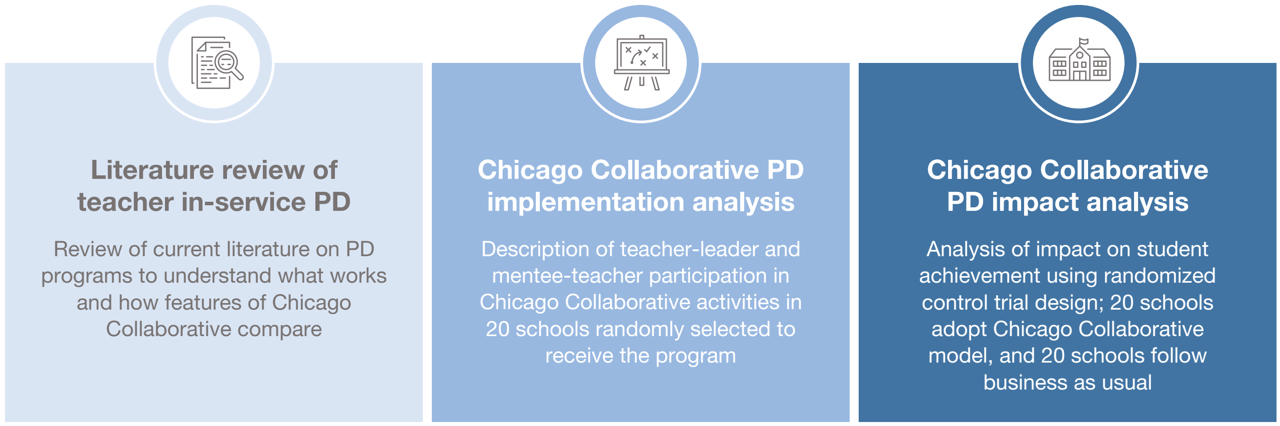
What Do We Know About the Impact of Teacher Professional Development?

To provide both a foundation and context for our analyses of the Chicago Collaborative PD program, we conducted a review of the existing literature on teacher in-service PD programs. This section summarizes the main findings of the first stage of our research.

Key features of PD programs with positive teacher and student outcomes: Meta-reviews as of this writing have identified the key features of teacher PD programs that are associated with positive teacher and student outcomes. The reviews distinguish between core features of programs, such as the content and focus of the program, and structural features of programs, such as the length and delivery. The core features of teacher PD programs associated with improved outcomes include having a content focus, targeting improvement in teacher pedagogical content knowledge, being evidence-based, and having designs in which teachers had ownership over the activities (Merchie et al., 2018). Structural features of programs that were associated with improved outcomes included being sustained over time, being embedded within teachers' daily work, being related to student learning and teaching standards, containing opportunities for peer collaboration, and engaging teachers afterward to reinforce material;

FIGURE 1

Chicago Collaborative Professional Development Program Research Methods in Brief



these programs show promise for improving student outcomes (Dunst, Bruder, and Hamby, 2015; Merchie et al., 2018; Wei, Darling-Hammond, and Adamson, 2010). Dunst, Bruder, and Hamby, 2015, highlighted the value of providing teachers with coaching, while Merchie et al., 2018, emphasized the importance of trainer quality and customized feedback. Common barriers to teacher PD implementation have also been identified, including limited school resources, time constraints, difficulties sustaining mandated curricula, and classroom management problems (Buczynski and Hansen, 2010). Sims and Fletcher-Wood, 2021, challenged consensus views on beneficial program features, including program duration, practice-based content, and peer collaboration, asserting that meta-reviews included studies that did not use rigorous methods to identify program effects.

PD and teacher-related outcomes: Prior research suggests that teacher PD can improve teacher-related outcomes, albeit with some mixed results. There is evidence that training programs can develop teachers' pedagogical content knowledge, including their ability to address students' content-related misconceptions (Morge, Toczek, and Chakroun, 2010) and implement inquiry-based practices (Buczynski and Hansen, 2010). However, Jacob, Hill, and Corey, 2017, observed little effect on teachers' mathematical content knowledge despite the intervention's extensive overlap with characteristics of effective PD (having a content focus, permitting teacher collaboration, and encouraging teacher ownership of the work). Other research has sought to identify instructional effects. Garet et al., 2008, evaluated a content-focused teacher PD series that occurred throughout the summer and much of the school year. They observed improvements in teachers' knowledge of evidence-based reading instruction and a greater likelihood of teachers using one of three instructional practices taught in the series.

PD and student-related outcomes: Some early studies found that teacher PD had no significant effects on student test scores (Harris and Sass, 2011; Jacob and Lefgren, 2004). However, a review by Yoon et al., 2007, estimated an increase of 21 percentile points on student achievement among nine studies that met What Works Clearinghouse Standards (and five of the studies met standards without reservations). Results were sensitive to the duration of the PD activities. The three studies that involved the least amount of programming (i.e., five to 14 hours total) showed no statistically significant effects. More-recent research (2019 to 2022) has continued to identify positive effects on student test scores, although the identified effects were small. In a meta-review of more than 50 teacher PD programs, Garrett et al., 2021, found that the average program effect on student achievement was 0.09 standard deviations (SDs), which was statistically significantly different from zero. Professional learning programs that facilitated teacher collaboration yielded larger average effects than those that did not (Lynch et al., 2019). In addition to student test scores, modest positive effects on students' social-emotional skills, including emotional comprehension and problem-solving skills, have been observed (Bierman et al., 2008).

PD and teacher coaching programs: In PD programs that offer this option, peers observe teachers' instruction and provide feedback. This kind of PD has emerged as a potential alternative to traditional PD interventions. Kraft, Blazar, and Hogan, 2018, conducted a meta-analysis of 60 studies that employed coaching. To be included, coaching needed to be one-on-one with teachers, occur at least every couple of weeks, be sustained over a long period of time, be relevant to the context of teachers' classrooms, and focus on deliberate skill development.[1] They found average effect sizes of 0.49 SD on instruction and 0.18 SD on achievement. Gains increased when, in addition to coaching, teachers participated in group training sessions (0.80 SD on instruction and 0.30 SD on achievement) or received instructional resources and materials (0.71 SD on instruction).

Although coaching programs are shown to have positive effects, it is unclear which specific features provide the necessary conditions for success. Blazar and Kraft, 2015, estimated the impact of a coaching program focused on improving teachers' pedagogical practices, examining how differences in program design were related to its effect on teacher cohorts. Teacher-to-coach ratios, coaching dosage, coach turnover, coach effectiveness, and content focus were all factors identified as playing a role in student achievement outcomes. However, the skill sets that characterized a high-quality coach were uncertain because neither the intensity of training nor coach workload was a driver of coach effectiveness.

Overview of Chicago Collaborative

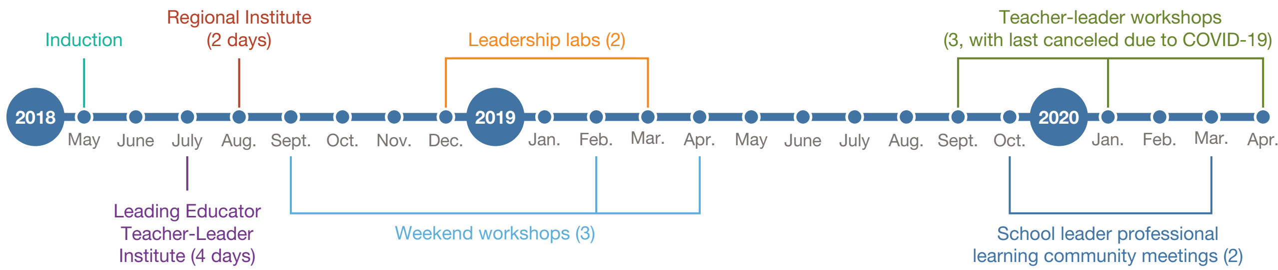
Leading Educators is a 501(c)(3) organization that partners with school districts to provide select teachers with PD, job-embedded coaching, and cohort-based learning experiences. The goal of the organization is to empower promising teacher leaders to raise the quality of instruction, improve culture schoolwide, and improve student outcomes. In 2016, Leading Educators received an Investing in Innovation (i3) grant from the U.S. Department of Education to launch the Chicago Collaborative, a cohort-based experiential training program aligned with CCSS. From January 2018 to March 2020, Leading Educators provided the two-year teacher PD program in 20 schools across three local district partners: Chicago Public Schools (CPS) and two local charter districts, anonymized here as District B and District C. The program trained one to three teacher-leaders from each participating school to deepen their pedagogical content knowledge in math and English language arts (ELA) and provided bimonthly, job-embedded coaching support to help them facilitate professional learning content cycles within their schools.

The Leading Educators Chicago team consisted of three leadership coaches, one program director, one program manager, two designers partially allocated to the program (to revise and refine professional learning sessions, content cycles, and learning labs), a senior director who managed the designers, and a managing director partially allocated to the program who led the collaboration with district leaders to customize the program. To inform program priorities and understand the contextual environment, the Leading Educators Chicago team completed three visits to schools during the 2018–2019 and 2019–2020 school years, where they observed two to five classrooms per school alongside school and teacher-leaders. The observations used Achieve the Core's Instructional Practice Guide, a resource that supports shifts in instructional practice required by the college- and career-ready standards, including the Common Core (Student Achievement Partners, undated). After the observations, the team reflected on what they saw and identified priorities for professional learning.

Teacher-leaders participated in intensive training activities throughout the two-year program (Figure 2). Starting in the spring of 2018, teacher-leaders attended a one-day induction meeting to begin formulating their identity as a regional cohort, build rapport as a team, and establish a safe learning culture. Next, they participated in a four-day intensive teacher-leadership institute focused on developing pedagogical content knowledge in math and ELA, introducing teacher-leaders to con-

FIGURE 2

Chicago Collaborative Teacher-Leader Training Activities



NOTE: COVID-19 = coronavirus disease 2019.

tent cycles, and deepening understanding of equity and antibias education. The final summer session was a two-day regional institute for school teams to collaborate and create plans for the upcoming school year and refine their content cycles. During the school year, teacher-leaders attended three one-day weekend workshops to develop skills to lead for equity and continued to build on cycles of professional learning. Finally, there were two leadership labs in the first year that allowed teacher-leaders to discuss best practices in leading adult learning. School leaders attended all of the same sessions as teacher-leaders, apart from the leadership labs. Full-day events were eight hours long, and leadership labs were three hours long.

Learning sessions continued in year two. Teacher-leaders attended two more workshops to further improve their facilitation strategies with their content cycle teams. In addition, school and district leaders participated in half-day professional learning communities to discover how to create the learning environments that content teams need to encourage growth in ELA and math, instructional practice, and analysis of student data.

Between learning sessions, teacher-leaders met with small groups of mentee teachers in their departments (one to five teachers) to facilitate three-week in-school professional learning content cycles with support from Leading Educators. Content cycle meetings occurred a few times each month and were embedded in the school day or held after school depending on teachers' schedules. Teacher-leaders noted that sessions occasionally needed to be postponed for school leaders to provide administrative updates. Prior to the start of a new content cycle, teacher-leaders analyzed student data and teacher practice data to identify gaps in teachers' pedagogical content knowledge as aligned with CCSS. With this information, they then planned cycles designed to increase their teams' abilities to deliver rigorous content and foster a culture of professional collaboration. Examples of content cycle topics included Building Knowledge and Vocabulary Through Literature, CCSS Math Shifts, Shifting the Lift in Literacy, Introduction to the Standards for Mathematical Practice, Close Reading with Complex Texts, and Adapting Learning for All Students.

Generally, content cycles ran in three-week patterns: shared learning (week one), planning and practice of that shared learning through co-planning (week two), and analysis of student data (week three). Teacher-leaders commonly adjusted the three-week pattern to spend more time on the content from week one. Mentee teachers learned about new material aligned with the topic of the content cycle, reviewed artifacts, evaluated and practiced lesson plans, and assessed student progress. Although Leading Educators' resources offered a starting point, teacher-leaders adjusted session plans to accommodate the localized needs of their schools. Teacher-leaders could also adjust the pace, commonly choosing to spend additional weeks on the shared learning component of the cycle. To better prepare their facilitation, teacher-leaders met bimonthly with Leading Educators leadership coaches to update them on the content cycle's progress and receive additional coaching. After administering surveys and completing school observations, leadership coaches completed reports for each school to help school and teacher-leaders monitor and analyze their own progress.

The Chicago Collaborative contains many effective features identified in the research literature (e.g., Dunst, Bruder, and Hamby, 2015; Merchie et al., 2018; and Wei, Darling-Hammond, and Adamson, 2010). First, the program's objective was to improve teachers' pedagogical knowledge in math and ELA with content-focused training aligned with teaching standards (i.e., the Common Core). Second, teacher-leaders participated in intensive training that was sustained over time. Over the course of two years, teacher-leaders completed up to 102 hours of required training activities, and school leaders completed up to 88 hours of training activities. Furthermore, teacher-leaders attended one-hour coaching sessions bimonthly throughout the program's duration. Third, content cycles for mentee teachers that were led by teacher-leaders provided opportunities for coaching and peer collaboration and allowed teachers to assume ownership over their professional learning. Content cycles were embedded in mentee teachers' regular work hours and extended the reach of the professional learning activities.

RAND's Evaluation of the Chicago Collaborative

The RAND Corporation's implementation and impact evaluation was conducted during the 2018–2019 and 2019–2020 school years in 40 schools located across three local districts in Chicago: CPS and two local charter districts. Given their smaller size, we have anonymized the two charter districts as District B and District C. Schools in CPS are grouped into local networks; the RAND research team worked with four of these networks and included eight schools from each local network. With assistance from Leading Educators, schools were recruited in summer 2017 to participate in the research. Prior to random assignment, the 40 school principals who consented to participate in the research selected teacher-leaders, such as grade-level subject chairs, and identified mentee teachers to be supported by teacher-leaders. In addition, principals selected the subject area (i.e., math or ELA) and grade-level range (i.e., grades 3 through 5, grades 6 through 8, or grades 3 through 8) that would be the focus of the intervention. Notably, this meant that the research team was able to identify the teachers and students in the schools that eventually served as control schools who would have been treated had the school been randomly assigned to treatment.

Random assignment was conducted in October 2017 using a rerandomization procedure described in the appendix. At the recruitment stage, districts participating in the research were asked to provide school-level administrative data about student achievement, student demographics, and scores on the 5Essentials school climate surveys. Schools were grouped into strata based on the local network, with four strata from CPS and one each from District B and District C, for a total of six strata. Schools were then randomized to the treatment or control condition within a stratum. All schools were contacted in November 2017 with their treatment assignment, and Leading Educators began working with school leaders to prepare for the delivery of the program in January 2018. Schools in the control condition operated under business-as-usual conditions, implementing teacher PD programs as they would in a typical school year.

The research team received implementation data from Leading Educators on teacher-leaders' participation in program activities, the frequency of coaching received by teacher-leaders from Leading Educators, and the frequency of content cycles overseen by teacher-leaders with their mentee teachers. We also obtained student-level administrative data from the three districts, which included demographic characteristics of students, such as gender, race, ethnicity, and free or reduced-price lunch (FRPL) status. Scores for the spring wave of the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) mathematics and ELA assessment were also obtained in each of the school years of 2016–2017, 2017–2018, and 2018–2019.[2] The students' subject-specific percentile growth from the 2017–2018 school year to the 2018–2019 school year served as the main outcome measure for the impact evaluation.[3]

The analytic sample thus consisted of all students in the grade band and subject combination named by the principal prior to random assignment who had prior test scores. In this analytic sample, we defined students' treatment status using the school they were enrolled in immediately prior to the randomization in November 2017.[4] The sample for the research is summarized in Table 1, which reports the demographic characteristics and baseline standardized test scores for students across the three districts. The sample is about 50 percent male, with a high fraction of minority students, and almost all students received FRPL.

How Was the Chicago Collaborative Implemented?

Leading Educators considered three key components of implementation: (1) school conditions, (2) program participation, and (3) teacher-leader interactions with mentee teachers.

TABLE 1

Sample Summary Statistics

Variable Name	CPS	District B	District C
Male	50.2%	48.6%	49.7%
Black	62.8%	88.0%	26.2%
Hispanic	35.9%	9.4%	67.4%
White	0.6%	0.2%	2.3%
Received FRPL	91.4%	100.0%	N/A
Student has IEP	16.4%	15.8%	13.2%
ELL student	0.0%	2.2%	42.0%
Math percentile score in 2018	38.0 (26.1)	56.9 (26.7)	25.4 (24.1)
ELA percentile score in 2018	44.3 (27.0)	54.4 (25.0)	29.4 (25.0)
Number of students	6,115	779	1,016
Number of schools	32	4	4

SOURCE: Based on authors' calculations using data provided by districts.

NOTE: ELL = English language learner; IEP = individualized education plan; N/A = not applicable. Sample includes all students in the estimation sample for the regressions. Districts B and C did not provide data on FRPL. However, District B noted that all students qualified for FRPL.

School conditions: Before the program began, school and district leaders assessed their schools' strengths and challenges according to six domains of school conditions that facilitate team-based, content-specific professional learning in schools. Leading Educators used this data to guide the implementation strategy and priorities. The domains were

- distributed instructional leadership (teachers are involved in improving instruction)
- diverse instructional leadership (leadership team is diverse across race, experience, etc.)
- aligned curriculum (embraces college and career readiness standards)
- aligned assessment (uses assessment data based in CCSS)
- master schedule (allocates time for teachers to participate)
- limited priorities (minimizes additional teacher responsibilities to ensure focus on planning and instruction).

Program participation: Leading Educators collected information about teacher-leaders' and school leaders' attendance at required training sessions (Table 2) and the frequency of teacher-leaders' attendance at bimonthly coaching sessions held by Leading Educators staff. Overall, program attendance rates were high. Most teacher-leaders and school leaders attended required training sessions. The average attendance rate of training activities among teacher-leaders was 84.9 percent across both years and in all three school districts. The average attendance rate was 79.5 percent among school leaders. Table 2 displays the average attendance rates by school districts and school year. Although participation among teacher-leaders declined in the second year, most of the training sessions occurred in year one, and therefore the decline in attendance in year two did not significantly decrease the overall attendance rate.

Teacher-leader interactions with leadership coaches: *There was variation in teacher-leaders' participation in coaching sessions with Leading Educators leadership coaches.* Given responses to a survey of teacher-leaders' engagement in coaching sessions, each school received a fidelity rating from Leading Educators, grouped into three categories: high

TABLE 2
Average Attendance Rates of Chicago Collaborative Training Activities

School District	2018–2019	2019–2020
All teacher-leaders	88.7%	67.8%
CPS	87.5%	79.3%
District B	85.7%	50.0%
District C	96.3%	91.7%
All school leaders	79.4%	80.0%
CPS	76.1%	79.5%
District B	92.9%	83.3%
District C	85.7%	62.5%

fidelity, medium fidelity, or low fidelity.[5] Across the 20 schools, nine received a rating of high fidelity (45 percent), four received a rating of medium fidelity (20 percent), five received a rating of low fidelity (25 percent), and two schools (10 percent) withdrew from the partnership after being assigned to the treatment group but before the program's launch.[6] Figure 3 shows the fidelity scores by school district.

Teacher-leader interactions with mentee teachers: Leading Educators surveyed leadership coaches to assess the interactions between teacher-leaders and mentee teachers at three points during the research period (fall 2018, spring 2019, and fall 2019). In these surveys, leadership coaches reported on the frequency of content-cycle meetings overseen by teacher-leaders, mentee teachers' attendance at scheduled meetings, the use of Leading Educators' materials, and barriers to implementation. The surveys suggest that *schools followed through on implementation to reach mentee teachers, but some variation existed*. Figure 4 shows the meeting frequency by district. In most schools, content cycles met a couple times a month or weekly. Throughout the program, fewer schools had leadership coaches who completed the survey, and all districts saw a decrease in meeting frequency over time. Among the responding teacher-leaders, CPS schools had the highest meeting frequency on average.

Finally, there is some additional implementation information collected in the survey about the content cycle meetings' quality. Across the three surveys, 95.7 percent of teacher-leaders reported that all or most mentee teachers were attending their scheduled content cycle meetings. Leading Educators' guiding materials were also widely used: 93.3 percent of teacher-leaders reported using the materials in some manner, and most either used the materials as written or made thoughtful modifications.

In fall 2018, Leading Educators reported little evidence that teachers were applying learning from the content cycles to their instructional planning and practice. Observations from the school

FIGURE 3
Teacher-Leader Coaching Fidelity

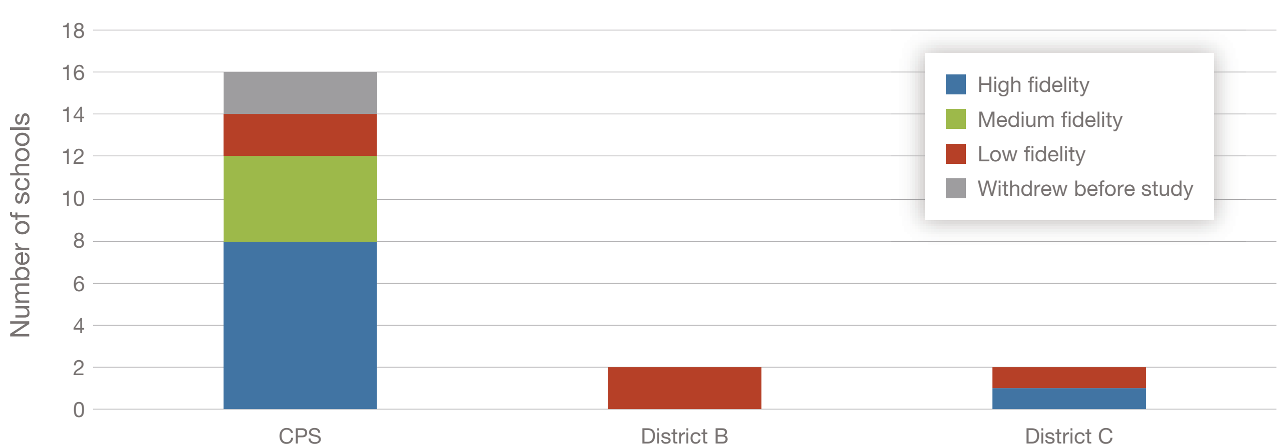
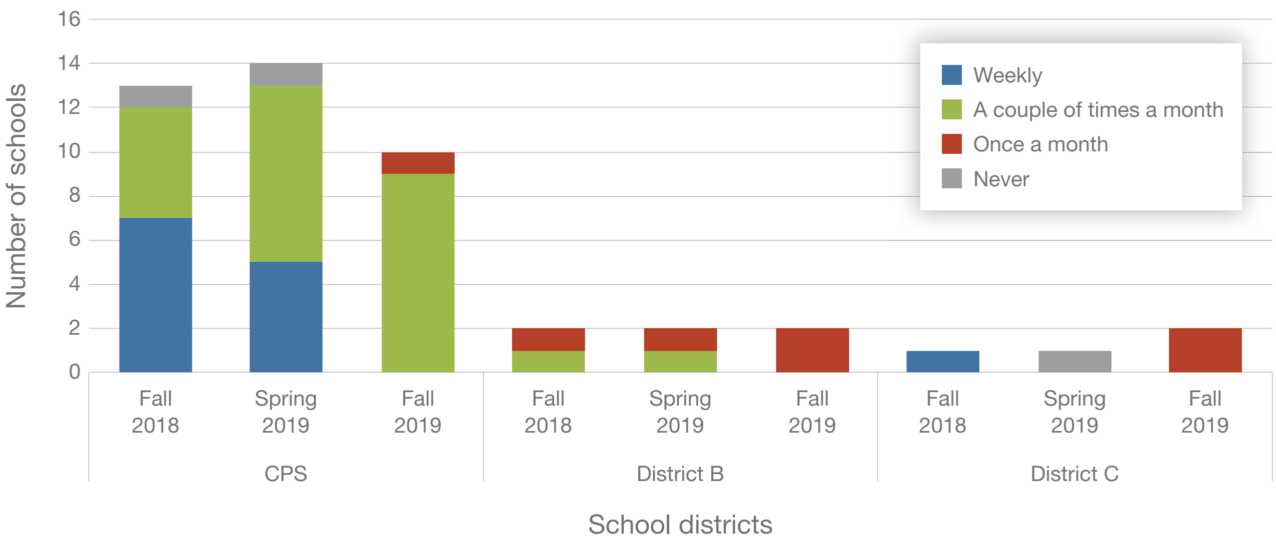


FIGURE 4
Content Cycle Meeting Frequency



NOTE: Responses were averaged across teacher-leaders within a school. Generally, teacher-leaders within schools reported the same meeting frequencies. In spring 2019, one school from CPS had an average meeting frequency between once a month and a couple of times a month, which was rounded up to a couple of times a month. The number of schools completing the survey differed at each checkpoint. *N* = 16 (fall 2018), 17 (spring 2019), and 14 (fall 2019).

visits suggested improved uptake of content cycle materials by the spring 2019 check-in, particularly in instructional planning, although some schools still experienced little evidence of change. Teacher leaders' facilitation of content cycle meetings also improved, and by fall 2019, most were reported to have strong leadership skills or had markedly improved throughout the program. Some common barriers to successful implementation of the content cycles that were listed by teacher-leaders were insufficient meeting time, inconsistent pacing of content cycle material, infrequent follow-ups between meetings to internalize materials, competing school priorities, difficulties aligning curriculum to standards, and difficulty translating content cycle content to instructional practice.

How Did the Chicago Collaborative Impact Student Outcomes?

We found robust evidence that the Chicago Collaborative increased student test scores. In our preferred specification, we compared the average test score growth in the targeted subject of students in schools randomly assigned to the treatment group with average test score growth in the subject that would have been targeted of students in schools randomly assigned to the control group, while including fixed effects for the schools' district or network.[7] We found that students in treated schools had higher test scores than students in control schools by 1.68 percentile points, which amounts to a difference of 0.05 SDs. The average for all schools is marked in the green column in Table 3. This estimated effect is statistically significantly different from zero at the 5-percent level and holds across a variety of specifications, such as when we controlled for a variety of student and school characteristics. Further validating the results, we found no impact on students' test scores in the nontargeted subject (see column five in Table 3). In other words, if the principals announced that their focus would be on math—as nearly all of the principals did—we found no impact on ELA test scores.

Although relatively small in magnitude, this effect is still meaningful, especially when considering the fact that teachers are only one driver of student test scores. When viewed from a teacher's

TABLE 3

Chicago Collaborative Impact on Student Achievement

	(1) Test Score Percentile Growth	(2) Test Score Percentile Growth	(3) Test Score Percentile Growth	(4) Test Score Percentile Growth	(5) Test Score Percentile Growth
Treatment effect	1.678**	1.719**	0.774	2.648**	−0.298
	(0.731)	(0.850)	(1.047)	(1.046)	(1.008)
Subject	Targeted subject	Targeted subject	Targeted subject	Targeted subject	Nontargeted subject
Covariates	None	Student demographics	None	None	None
Population	All	All	Elementary schools	Middle schools	All
Number of clusters	251	246	177	184	249
Number of students	7,911	6,891	2,748	4,252	7,708

NOTE: ** indicates significant at $p < 0.05$. Green shading indicates the average for all schools.

perspective, for example, the effect moves a teacher from being an average teacher to one who is better than two-thirds of teacher peers at increasing student test scores.[8]

Given the small school-level sample size, we were unable to explore in detail whether the program had differential effects across subgroups and whether the program had effects on other outcomes. We found suggestive evidence that the Chicago Collaborative had a larger impact in middle schools than in elementary schools. However, we cannot reject (at the 5-percent level) the null hypothesis that average effects in the two subgroups are identical. Similarly, although we found no evidence that the Chicago Collaborative impacted the 5Essentials school climate measures, the low school-level sample size meant that the estimates were quite uncertain. We likewise cannot rule out the possibility that the program had a meaningful impact on school climate. We can rule out the possibility that the program increased or decreased student attendance by more than 0.5 percentage points.[9]

Limitations of This Research

Because of the COVID-19 pandemic, the implementation of the program was cut short, and the research was only able to analyze the first year of implementation of the two-year PD program. Specifically, spring 2020 test scores would have served as the primary outcome for analysis, but because of school closures in March 2020, these test scores were never collected. Leading Educators' theory of change suggests that our estimate of the program's impact on student achievement is conservative. This is because it takes time for teacher-leaders to effectively transfer the knowledge and skills gained in training to their own practice, additional time for teacher leaders to pass this knowledge to their peers, and even more time for mentee teachers to transfer this knowledge into improved outcomes for their students.

Another limitation of the research is that additional outcomes of interest were not available for analysis. First, we were unable to analyze the impact of the program on such important teacher outcomes as retention because of the shortened time frame for data collection. Also, although we had intended to analyze teacher pedagogical content knowledge through the use of valid and reliable scales collected in online surveys, our survey response rates were too low to be included in the research. Lastly, the research had intended to analyze PARCC test scores that are designed to assess CCSS. Because these scores were not available, the research used NWEA MAP test scores instead. Although there is evidence that the test scores in our model are aligned to CCSS for a different state, they were not

Our research found that the Leading Educators Chicago Collaborative was implemented overall as intended and had a statistically significant positive effect on student achievement.

the intended outcome for an intervention that is focused on CCSS in Illinois (Set, 2018).

Summary and Next Steps

Our research found that the Leading Educators Chicago Collaborative was implemented overall as intended and had a statistically significant positive effect on student achievement. The variation we observed in implementation was what could be expected were the program to be scaled to a larger set of schools (with relatively high attendance in training sessions and variation in leadership coaching and content cycle frequency). The positive and statistically significant effect of the program on student achievement was observed despite the fact that the assessment we analyzed was not specifically aligned to CCSS, which was a focus for the intervention, nor were we able to fully capture the effects of the two-year program (because we were missing test scores from spring 2020). We were able to document suggestive evidence of differential effects of the program on some subgroups, but this analysis was hampered by small sample sizes at the level of treatment (the school). Given the positive findings, additional evaluations are needed to understand which features of the program are effective and whether they are effective for subgroups of schools or students.

APPENDIX
Methods

Data cleaning and merging: CPS and Districts B and C provided administrative data for four school years: 2016–2017 through 2019–2020. The data included the NWEA MAP test scores for all tested grade levels. Because of the program's attention to the Common Core, we would have preferred PARCC test scores or an equivalent Common Core–aligned assessment. However, the project was not able to obtain these data. Furthermore, test scores were unavailable for spring 2020, the end of the second year of program implementation, because of lockdowns associated with the COVID-19 pandemic. We also received student-level demographics (e.g., age, gender, race, IEP, FRPL, and ELL), attendance, and discipline data. We cleaned these administrative data to facilitate merging across all districts and creating uniform demographic indicators. We also identified students who had transferred schools or repeated grades.

To merge the data, we first defined our sample only to include students attending the schools involved in this research during the randomization school year, 2017–2018. We combined these data with the other school years via student identification cards, resulting in data sets for each school from 2016–2017 through 2019–2020 of only active students in the randomization school year. Next, we merged the three districts. Finally, we merged these data with the school-level treatment assignment and 5Essentials survey data of school climate. Note that the 5Essentials survey data is measured at the school level, and unlike the rest of our analysis, this information necessarily includes joiners. Our analytic sample therefore does not include students who joined schools after random assignment.

Rerandomization procedure: We assigned treatment to schools using a rerandomization procedure (Morgan and Rubin, 2012). In this procedure, we first divided schools into six strata (four local networks in the three districts). To measure the difference in characteristics between the two groups (within strata), we calculated the *Mahalanobis distance*, a measure of group difference (Tanton, 2005), between the treated and control schools. The covariates used in the Mahalanobis distance calculation were school-level scores from the 2016 5Essentials school climate survey (domains were effective leaders, collaborative teachers, supportive environment, ambitious instruction, and involved families) and school-level scores from the spring 2016 PARCC tests (mean standardized math score, mean standardized ELA score, percentage of students who are proficient in math, and percentage of students who are proficient in ELA).

Prior to the assignment of schools to treatment and control groups, we defined a strict cutoff for an acceptance difference in the Mahalanobis distance as 0.5. This meant that if the Mahalanobis distance was more than 0.5, we rejected the treatment assignment and repeated the random assignment of schools to treatment (within strata) and recomputed the Mahalanobis distance. We repeated this procedure until we found a random assignment where the Mahalanobis distance was less than 0.5. This cutoff meant that the procedure would accept only one out of every 30,000 randomization attempts, ensuring very strong balance.

The fact that we eventually found a solution indicates that there was a treatment assignment with a Mahalanobis baseline balance of less than 0.5. Because exactly half of the schools were treated, an assignment in which all the treated schools were instead assigned to the control and all the control schools assigned to treatment would also have a Mahalanobis baseline balance of less than 0.5. As is clear from our randomization inference procedure described next, there were many assignments that have a Mahalanobis baseline balance of less than 0.5, indicating that our assignment was not deterministic, and the cutoff was not set to be too stringent.

Balance between treatment and control groups: The rerandomization procedure was done to ensure that we had balance across treatment and control groups on the 5Essentials measures and PARCC test scores. We now show that the rerandomization procedure balanced not only these outcomes but also a handful of other measures. We start by showing balance on the measures which are only available at the school level, displayed in Table A.1. Note that the number of schools is less than 20 in each column due to the fact that we did not have school-level baseline measures for District C schools.

TABLE A.1
Baseline Equivalence, School Level

		Treatment Condition			Control Condition		
Category	Measure	Number of Schools	Mean	Standard Deviations	Number of Schools	Mean	Standard Deviations
5Essentials	Ambitious instruction	17	70.10	17.60	17	69.50	16.20
5Essentials	Effective leadership	17	51.90	18.60	17	50.20	14.80
5Essentials	Collaborative teachers	17	58.60	18.60	17	58.20	17.20
5Essentials	Involved families	17	58.50	15.90	17	59.20	14.60
5Essentials	Supportive environment	17	49.90	15.20	17	47.10	22.40
PARCC test scores	Average math scores	17	–0.51	0.49	17	–0.53	0.40
PARCC test scores	Average ELA scores	17	–0.45	0.49	17	–0.46	0.46
PARCC test scores	Percentage of students who are proficient in math	17	0.34	0.14	17	0.36	0.09
PARCC test scores	Percentage of students who are proficient in ELA	17	0.43	0.14	17	0.45	0.11
School size	Number of full-time employees	18	30.40	12.00	18	28.90	14.80
School size	Number of students	18	511.00	219.00	18	512.00	335.00

We next show the balance of characteristics at the student level, focusing only on students in the analytic sample. This is displayed in Table A.2.

Methods to adjust for student demographics: To estimate the effect of the treatment, we estimated Ordinary Least Squares regressions of the form

$$y_{it} = \alpha_{strata} + \tau T_{s(i,t-1)} + \beta X_{it} + \in_{it}$$

where y_{it} is the outcome of interest for individual i in year t, $T_{s(i,t-1)}$ is a dummy variable that indicates whether individual i was attending a treated school in the previous year when the random assignment was done, and X_{it} are a set of student i's demographics.

In all specifications, we included strata fixed effects (α_{strata}) and clustered the error term ($\in_{it}$) at the school level. Our preferred specification did not include any student covariates X_{it} because doing so slightly lowered the sample size. In the one specification where we did include covariates, we included the students' grade, gender, race, and ethnicity; whether they qualify for FRPL; whether they are classified as an ELL; and baseline test score percentile growth.

Imperfect compliance: In our main analysis, we conducted an intent-to-treat estimate, in which we considered any individual treated if they were in a school that was randomly assigned regardless of whether they were exposed to the treatment. We also estimated a treatment-on-the-treated (TOT) effect in which we used treatment assignment from the rerandomization procedure as an instrument for whether students received treatment or not. In this context, there were two reasons why students might not have received the treatment to which they were randomly assigned. First, some students might have moved from schools assigned to the treatment group to other schools—either schools that were assigned to the control group or schools that were not in the

TABLE A.2

Baseline Equivalence, Student Level

Measure	Treatment Condition				Control Condition			
	Number of Individuals	Number of Schools	Mean	Standard Deviations	Number of Individuals	Number of Schools	Mean	Standard Deviations
Black	4,003	20	0.63	0.48	3,764	20	0.60	0.49
Hispanic	4,003	20	0.35	0.48	3,764	20	0.38	0.49
Male	4,139	20	0.50	0.50	3,769	20	0.50	0.50
ELL	4,004	20	0.04	0.20	3,764	20	0.06	0.23
FRPL	3,585	16	0.92	0.27	3,309	16	0.93	0.25
Homeless	3,608	16	0.05	0.20	3,381	16	0.05	0.22
Repeat grade	4,142	20	0.13	0.12	3,769	20	0.12	0.11
Test score percentile growth	3,527	20	–1.40	16.70	3,296	20	–1.15	17.60

research at all.[10] Second, two of the schools assigned to the treatment status (Kipling and Hendricks) did not implement the program.[11]

To estimate the TOT, we ran a two-stage least squares regression, using the treatment assignment as an instrument for treatment receipt. Doing so increased the estimated impact of the treatment from 1.68 percentile points—reported in column (1) of Table 3—to 2.12 percentile points, with a standard error of 0.95 and a p-value of 0.03.

Attrition: Because we used administrative data, we had very low attrition. Specifically, we did not have any attrition at the school level. The only attrition from the analytic sample occurred when students left the district between the time of randomization and the time of the analysis. The level of attrition is shown in Table A.3.[12] The overall student attrition rate was 11.0 percent, and the differential student attrition rate was 1.1 percent.

Randomization inference: In addition to testing the null hypothesis that the program had no effect using traditionally estimated standard errors and p-values, we also used randomization inference to test the null hypothesis without employing asymptotic analysis. In this procedure, we repeated the random assignment of schools to treatment and control groups using the same process described in the "Rerandomization Procedure" section. We ran this procedure numerous times, leading to many other potential treatment assignments. For terminology, we called the schools assigned to treatment the *placebo-treated schools* and the schools assigned to control the *placebo-control schools* because they were not the realized treatment assignment. We then compared the average outcome of the placebo-treated schools with the average outcome of the placebo-control schools for each of these other potential treatment assignments. This gives rise to a distribution of potential effect estimates under the null hypothesis of no treatment effect.

When we compared our estimated effect with the distribution potential effect estimates under the null hypothesis, we found that only 5.7 percent of the distribution consisted of effects with a larger absolute magnitude than the estimated effect distribution. By definition, this corresponds to a p-value of 0.057, which is similar to the p-value we found using the traditional methods.

Other outcomes: In Table A.4, we report the point estimates, standard errors, and sample size for the other outcomes we measured. As mentioned in the main body of the report, we found a relatively precise zero effect on attendance rates and noisy estimates on all of the 5Essentials survey outcomes.

TABLE A.3

Attrition at the School and Student Level

Count	Treatment Condition	Control Condition
Number of clusters randomly assigned	20	20
Number of clusters in the analytic sample	20	20
Number of students in clusters before random assignment	4,684	4,208
Number of students in analytic sample	4,142	3,769

TABLE A.4

Chicago Collaborative Impact on Attendance and 5Essentials Survey Measures

	(1) Attendance Rate Growth	(2) Effective Leaders Growth	(3) Collaborative Teachers Growth	(4) Involved Families Growth	(5) Supportive Environment Growth	(6) Ambitious Instruction Growth
Treatment effect	–0.0028	0.58	–1.58	4.82	1.38	6.70
	(0.0018)	(4.65)	(4.68)	(4.95)	(7.02)	(8.00)
Covariates	None	None	None	None	None	None
Population	All	All	All	All	All	All
Number of clusters	251	N/A	N/A	N/A	N/A	N/A
Number of students	7,924	33	33	33	33	33

NOTE: N/A = not applicable.

Notes

[1] The study excluded teacher preparation and school-based teacher induction programs and coaching programs directly engaging with students.

[2] We also received the winter wave of NWEA MAP scores for two of the three districts (CPS and District C). The estimated effects that are reported next are similar when using the winter wave of test scores as the outcome and restricting the sample to CPS and District C.

[3] The MAP scores used in the analysis have a test-retest reliability of 0.72 to 0.87 and a predictive validity of 0.57 to 0.73 (Wang, Jiao, and Zhang, 2013). The study team had planned to use the Partnership for Assessment of Readiness for College and Careers (PARCC) assessment for the outcome analysis. However, in 2019, the PARCC was replaced with the Illinois Assessment of Readiness. Then, because of the COVID-19 pandemic, there was no testing in 2020.

[4] Thus, the only students who attrited from the study were ones that left the district. Our analytic approach also means that resulting estimates are intent-to-treat estimates. We discuss attrition and adjustments for imperfect compliance in the appendix. We also included in our analytic sample the two schools that were assigned to the treatment group but never implemented the program.

[5] To receive a score of *high fidelity*, coaching must occur at least bimonthly throughout the year. *Medium fidelity* indicated that coaching consistency varied, meaning that teacher-leaders were not present at all sessions or a few coaching sessions were canceled entirely. *Low fidelity* meant that coaching was sporadic, and teacher-leaders missed several months of sessions.

[6] As discussed next, the two schools that withdrew were still included in the analysis and so did not officially attrit from the study.

[7] Because all the schools in the same strata chose the same subject, these fixed effects also serve as subject fixed effects.

[8] This calculation uses the fact that a one-standard-deviation increase in teacher quality, as measured by their value-added scores, increases student test scores by approximately 0.2 SDs (Mulhern and Opper, 2022).

[9] The appendix presents the results from the analysis of the program impact on school climate and student attendance.

[10] In theory, students could also move from the control schools to the treated schools. In practice, that never happened. Although students moved into the treated schools from outside the study, we omitted these joiners from the analysis. Similarly, all of the students who moved from treated schools to other schools during the time frame moved to schools outside the study rather than to control schools inside the study. We considered those moving to control schools for this analysis.

[11] Two other schools (Burke and Colemon) stopped implementation after the first year, but we consider them as fully treated for this analysis because most of the training received by teacher-leaders occurred in year one of the program.

[12] To calculate the number of students in clusters before random assignment, we counted students in the schools but omitted those in grade 8, who would have naturally graduated out of the sample by the time of the analysis.

References

Bierman, Karen L., Celene E. Domitrovich, Robert L. Nix, Scott D. Gest, Janet A. Welsh, Mark T. Greenberg, Clancy Blair, Keith E. Nelson, and Sukhdeep Gill, "Promoting Academic and Social-Emotional School Readiness: The Head Start REDI Program," *Child Development*, Vol. 79, No. 6, November–December 2008, pp. 1802–1817.

Blazar, David, and Matthew A. Kraft, "Exploring Mechanisms of Effective Teacher Coaching: A Tale of Two Cohorts from a Randomized Experiment," *Educational Evaluation and Policy Analysis*, Vol. 37, No. 4, December 2015, pp. 542–566.

Buczynski, Sandy, and C. Bobbi Hansen, "Impact of Professional Development on Teacher Practice: Uncovering Connections," *Teaching and Teacher Education*, Vol. 26, No. 3, April 2010, pp. 599–607.

Dunst, Carl J., Mary Beth Bruder, and Deborah W. Hamby, "Metasynthesis of In-Service Professional Development Research: Features Associated with Positive Educator and Student Outcomes," *Educational Research and Reviews*, Vol. 10, No. 12, June 2015, pp. 1731–1744.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, Laura Sztejnberg, and Marsha Silverberg, *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, NCEE 2008-4030, September 2008.

Garrett, Rachel, Qi Zhang, Martyna Citkowicz, and Lauren Burr, *How Learning Forward's Standards for Professional Learning Are Associated with Teacher Instruction and Student Achievement: A Meta-Analysis*, Arlington, Va.: Center on Great Teachers and Leaders at the American Institutes for Research, December 2021.

Harris, Douglas N., and Tim R. Sass, "Teacher Training, Teacher Quality and Student Achievement," *Journal of Public Economics*, Vol. 95, No. 7–8, 2011, pp. 798–812.

Hill, Heather C., "NEPC Review: *The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development*," Boulder, Colo.: National Education Policy Center, September 2015.

Jacob, Brian A., and Lars Lefgren, "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago," *Journal of Human Resources*, Vol. 39, No. 1, Winter 2004, pp. 50–79.

Jacob, Robin, Heather Hill, and Doug Corey, "The Impact of a Professional Development Program on Teachers' Mathematical Knowledge for Teaching, Instruction, and Student Achievement," *Journal of Research on Educational Effectiveness*, Vol. 10, No. 2, 2017, pp. 379–407.

Kraft, Matthew A., David Blazar, and Dylan Hogan, "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence," *Review of Educational Research*, Vol. 88, No. 4, 2018, pp. 547–588.

Lynch, Kathleen, Heather C. Hill, Kathryn E. Gonzalez, and Cynthia Pollard, "Strengthening the Research Base That Informs STEM Instructional Improvement Efforts: A Meta-Analysis," *Educational Evaluation and Policy Analysis*, Vol. 41, No. 3, September 2019, pp. 260–293.

Merchie, Emmelien, Melissa Tuytens, Geert Devos, and Ruben Vanderlinde, "Evaluating Teachers' Professional Development Initiatives: Towards an Extended Evaluative Framework," *Research Papers in Education*, Vol. 33, No. 2, 2018, pp. 143–168.

Mihaly, Kata, Benjamin K. Master, and Cate Yoon, *Examining the Early Impacts of the Leading Educators Fellowship on Student Achievement and Teacher Retention*, Santa Monica, Calif.: RAND Corporation, RR-1225-LED, 2015. As of May 4, 2022: https://www.rand.org/pubs/research_reports/RR1225.html

Morgan, Kari Lock, and Donald B. Rubin, "Rerandomization to Improve Covariate Balance in Experiments," *Annals of Statistics*, Vol. 40, No. 2, April 2012, pp. 1263–1282.

Morge, Ludovic, Marie-Christine Toczek, and Nadia Chakroun, "A Training Programme on Managing Science Class Interactions: Its Impact on Teachers' Practices and on Their Pupils' Achievement," *Teaching and Teacher Education*, Vol. 26, No. 3, 2010, pp. 415–426.

Mulhern, Christine, and Isaac M. Opper, "Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness," Providence, R.I.: Annenberg Institute at Brown University, EdWorkingPaper No. 21-451, April 11, 2022.

The New Teacher Project, *The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development*, New York, 2015.

Sellen, Peter, *Teacher Workload and Professional Development in England's Secondary Schools: Insights from TALIS*, London: Education Policy Institute, October 2016.

Set, Andy, "Study Concludes MAP Growth Items Align to Common Core State Standards," Northwest Evaluation Association, February 27, 2018.

Sims, Sam, and Harry Fletcher-Wood, "Identifying the Characteristics of Effective Teacher Professional Development: A Critical Review," *School Effectiveness and School Improvement*, Vol. 32, No. 1, 2021, pp. 47–63.

Student Achievement Partners, "Instructional Practice Guide," webpage, undated. As of June 10, 2022: https://achievethecore.org/page/1119/instructional-practice-guide

Tanton, James, *Encyclopedia of Mathematics*, New York: Infobase, 2005.

Wang, Shudong, Hong Jiao, and Liru Zhang, "Validation of Longitudinal Achievement Constructs of Vertically Scaled Computerised Adaptive Tests: A Multiple-Indicator, Latent-Growth Modelling Approach," *International Journal of Quantitative Research in Education*, Vol. 1, No. 4, 2013, pp. 383–407.

Wei, Ruth Chung, Linda Darling-Hammond, and Frank Adamson, *Professional Development in the United States: Trends and Challenges*, Dallas, Tex.: National Staff Development Council, July 2010.

Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley, *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest, Institute of Education Sciences, U.S. Department of Education, REL 2007–No. 033, October 2007.

Acknowledgments

We are extremely grateful to Leading Educators, and especially Rebecca Taylor-Perryman, Shyla Kinhal, and Debbie Sim, for their support of the research. Their guidance and assistance in managing the project was critical. We are grateful to district staff for answering our requests and providing the data. Monica Mean provided invaluable project management support. We also would like to thank Christopher Doss and Andrew McEachin for helpful feedback that greatly improved this report and Monette Velasco for overseeing the publication process.

About This Report

The authors evaluated the Chicago Collaborative, a teacher professional development program that is aligned to Common Core State Standards and implemented by Leading Educators, a national nonprofit organization that partners with districts and charter management organizations to help teachers develop the leadership skills that they need to successfully transition from leading students to leading their peers. The authors conducted a randomized control trial evaluation using data from 40 schools across three school districts in the Chicago area during the 2018–2019 and 2019–2020 school years. They examined how the Chicago Collaborative program was implemented and whether the program impacted student achievement. The authors found that the Chicago Collaborative was successfully delivered, despite the challenges posed by the coronavirus pandemic at the end of the research period in 2020. The authors also found robust evidence that the Chicago Collaborative increased student test scores.

RAND Education and Labor

This study was undertaken by RAND Education and Labor, a division of the RAND Corporation that conducts research on early childhood through postsecondary education programs, workforce development, and programs and policies affecting workers, entrepreneurship, and financial literacy and decisionmaking. This study was sponsored through funding from the U.S. Department of Education Investing in Innovation (i3) grant (grant number U411C160040). The findings and conclusions presented are those of the authors and do not necessarily reflect positions or policies of the U.S. Department of Education.

More information about RAND can be found at www.rand.org. Questions about this report should be directed to kmihaly@rand.org, and questions about RAND Education and Labor should be directed to educationandlabor@rand.org.

For more information on this publication, visit www.rand.org/t/RRA2047-1.

www.rand.org

RR-A2047-1