Research Report

Examining the Early Impacts of the Leading Educators Fellowship on Student Achievement and Teacher Retention

Kata Mihaly, Benjamin K. Master, Cate Yoon

RAND CORPORATION

For more information on this publication, visit www.rand.org/t/rr1225

Published by the RAND Corporation, Santa Monica, Calif.

Preface

Leading Educators is dedicated to training promising mid-career teachers to retain effective educators and raise student achievement. RAND is conducting an independent multiyear evaluation of the Leading Educators Fellowship program. This report presents findings from the preliminary analyses of the program's effects. This report includes an analysis of program impacts on student achievement and teacher retention in both Louisiana and Missouri in the 2011–2012 through 2013–2014 school years for teachers who are program fellows and for the teachers mentored by fellows.

The findings from this study will be of interest to those concerned with programs and policies that provide support to teachers and teacher leaders and, more generally, programs that improve student achievement and teacher retention.

Support for this research was provided by Carnegie Corporation of New York, Charles and Lynn Schusterman Family Foundation, and Walton Family Foundation.

This research has been conducted in RAND Education, a unit of the RAND Corporation. Additional information about the RAND Corporation is available at www.rand.org.

Contents

Figures

Tables

Summary

This report is the first in a multiyear evaluation of the impact of the Leading Educators (LE) Fellowship program on participating teachers. Founded in 2008, the LE Fellowship program selects promising mid-career teachers through a competitive application process and develops their skills as instructional mentors and leaders of school improvement efforts. The specific objectives of the program are to (1) increase the leadership skills and capacity of teacher leaders in order to improve student achievement in high-need schools and (2) retain highly effective teachers in high-need schools by encouraging commitment to the schools and improving the school leadership pipeline.

During the course of the two-year program, fellows participate in a number of activities to improve their leadership skills. These activities include a one-week summer intensive training session, one-on-one leadership coaching, professional development sessions, and site visits to exemplary high-need schools. In addition, fellows select teachers within their school to mentor, depending on their leadership role in the school or needs identified by the principal or other district administrators. Starting in the fall of their first year, fellows implement their leadership skills by supporting mentee teachers to improve their students' performance. The mentoring activities led by fellows depend on their leadership role in the school and include coaching, observation and feedback, grade-level meetings, and problem-solving communities.

This report presents the preliminary findings of the evaluation, including a descriptive summary of teachers' skill development as measured by the program and a more rigorous analysis of the impact of the LE Fellowship program on student achievement and teacher retention in Missouri and Louisiana in the 2011–2012 through 2013–2014 school years. In support of the impact analysis, we used data from the LE Fellowship program, including application information and mentee teacher assignments, and combined it with state administrative data on students and teachers in both states. We examined the impact of the fellowship program on student achievement and teacher retention for fellows as well as the teachers they mentor.

Due to the small sample of teachers who were matched to the state administrative data panels, we include in our impact analysis of fellows not only program graduates but also fellows who are still midway through the program. Similarly, mentee teachers who are assigned to first-year fellows are included in the analysis. Future analyses will include a larger sample of participating teachers and may also allow for a more nuanced investigation of program effects, both during and subsequent to participation in the program.

We employed a variety of rigorous methodologies to compare student achievement gains for teachers who participated in the program as fellows or as mentees with the student achievement

gains of other teachers. In particular, we used three complementary methods for estimating program impacts:

- First, we compared fellows with other program applicants who were not accepted into the program, using controls for the application scores that determined whether or not they were accepted into the program and that may help to account for any differences between fellows and other applicants.
- Second, for both fellows and mentees, we utilized a propensity score matching technique in which we compared outcomes between participating teachers and comparison teachers who appeared similarly likely to have been program participants, based on their characteristics and the characteristics of the students that they teach.
- Third, for both fellows and mentees, we compared outcomes between participating teachers and a sample of comparison teachers who had similar instructional effectiveness in the year prior to the program's start.

Each of our three approaches has both unique advantages and limitations, and collectively they provide a robust investigation of possible early program effects than anyone method individually.

In addition to evaluating program effects on participating teachers' student achievement, we explored the impact of the program on teacher retention rates among both fellows and mentee teachers, and considered retention in the same school, retention in any high-poverty school in the state, and retention in any teaching position in the same state. For mentee teachers, where sample sizes permitted, we utilized the propensity score matching technique to compare teacher retention of mentees to comparison teachers who appeared similarly likely to have been program participants.

The findings from this early analysis of the program impacts of fellows are promising but mixed, and overall do not conclusively demonstrate that the program has affected student achievement. Among fellows, we find both some statistically significant positive and negative program effects on student achievement, with results that vary across states, subject areas, and model specifications. Only one finding, a positive and statistically significant effect on mathematics achievement in Louisiana, was confirmed by all three modeling specifications. In general, the estimates of student achievement effect for fellows are based on very small samples and do not point to any clear positive or negative impacts, so we recommend caution in interpreting these results. It may be that in some cases the fellowship training and interactions with mentees improved instructional effectiveness for some fellows. In other cases, it is possible that the additional responsibilities involved in program participation coupled with limited release time reduced fellows' instructional effectiveness. The inconsistency of our results for fellows across modeling specifications may also indicate that fellows are unique in unobserved ways that make it difficult to identify a truly fair comparison group.

Among mentee teachers, for whom our sample sizes are larger, we find some suggestive evidence of program impacts. In particular, we find marginally significant and significant positive program effects among mentees who teach math and social studies, respectively, in

Louisiana, and these results are fairly consistent across modeling specifications. However, we find no significant program effects in reading or mathematics in Missouri, nor in Louisiana in reading or science. In addition, an adjustment for multiple hypothesis testing would eliminate the statistical significance of both positive program effects observed in Louisiana. Thus, while our results indicate that it is possible that mentees in Louisiana are benefiting from their exposure to the program, additional evidence is needed to confirm this pattern.

Finally, the impact of the program on teacher retention is unclear, with no consistent pattern of retention impacts across cohorts or states. In some state-year cohorts, mentees in the program were significantly more likely to remain in the same school or in teaching. However, in other cases mentees were significantly less likely to remain in their school or in teaching. It is likely that we are unable to appropriately account for all external sources of variation in teacher retention in our analyses. Future analyses with larger samples of both fellows and mentees will provide greater statistical power for more reliably detecting any true program effects.

Acknowledgments

We thank Seth Stahlberger, Cate Yoon, and Tiffany Tsai for analysis support and Scott Naftel for programming support. We thank Rebecca Taylor-Perryman at Leading Educators for providing the data and assisting with the data merging. Finally, we thank Francisco Perez-Arce and our external reviewer for valuable feedback on all aspects of the report, including the presentation and empirical methods, and Catherine Augustine for comments on the manuscript.

1. Introduction

The Leading Educators (LE) Fellowship program selects promising mid-career teachers and develops their skills as instructional mentors and leaders of school improvement efforts. The LE fellows examined in this report participated in a two-year training program that consisted of a series of professional development sessions, school visits, and meetings with a leadership coach. Each year, fellows select teachers to mentor based on the leadership role that they play in their school and the needs identified by the principal or other district administrators. Mentoring meetings between fellows and their mentees occur frequently throughout the school year. Fellows are required to demonstrate mastery of leadership core modules, and they are evaluated on the Leadership Competency Rubric multiple times during the LE Fellowship program. The ultimate goal of the LE Fellowship program is to improve school outcomes by increasing student achievement and encouraging high-quality teachers to remain in high-poverty schools.

This report presents the preliminary findings on the impact of the LE Fellowship program on student achievement and teacher retention in Missouri and Louisiana in the 2011–2012 through 2013–2014 school years. We used data from the LE Fellowship program, including application information and mentee teacher assignments, and state administrative data on students and teachers in both states to perform the analysis. We employed a variety of rigorous analytical methods to measure the differential impacts of teachers influenced by the LE Fellowship program versus comparable teachers. Because leader development efforts may influence students of fellow and mentee teachers, we examined achievement outcomes across students taught by both populations. We also explored the impact of the program on teacher retention rates among both fellows and mentee teachers, including retention in the same school, retention in high-poverty schools in the state, and retention in teaching in the state.

In the remainder of this report, we first describe the background literature on teacher coaching and professional development programs that include components similar to those of the LE Fellowship program. Next, we describe the LE Fellowship program in detail. Following this, we discuss the data used for the student achievement and retention analyses in Missouri and Louisiana, including the sample of fellows included in the analysis and the demographic characteristics of participating teachers and of the students they teach. This section also includes descriptive analysis of the Leadership Competency Rubric, an observation protocol that was used to evaluate leadership competencies during the course of the fellowship. We then discuss the methods used to perform our evaluation of impacts, and present our results. We conclude with a discussion of the study's implications.

2. Background

At this time, there is little evidence on the impact of teacher leadership development programs. However, extant research on teacher development includes studies evaluating a range of in-service programs and interventions that partially overlap with components of the LE Fellowship program. In particular, research on coaching and formal and informal mentoring programs may relate both to the coaching and mentoring that fellows received and to the mentoring that the fellows provided to mentees in their schools. Similarly, research on professional development interventions may relate to the professional development that fellows received as a core component of their program experience. Overall, research on teacher development interventions has shown mixed results, but also indicates the potential for substantial benefits to students and teachers in some cases.

Recent research indicates that even in the absence of any formalized leadership program or intervention, teachers can have substantial impacts on their peers' assessed performance, with spillover effects on teacher productivity documented in middle and high schools (Yuan, 2015; Koedel, 2009). Moreover, Jackson and Bruegmann (2009) observed that teachers, particularly less experienced teachers, learn from peers who teach in the same grade level and subject area, as evidenced by apparent learning effects that persist over time. Overall, they found that historical peer quality explains between 18 and 25 percent of observed own-teacher effects.

In addition to evidence of teachers' informal impacts on peer performance, a growing body of research has examined the potential impacts of more structured interventions designed to spur peer-to-peer teacher development, including teacher mentoring, coaching, and new-teacher induction interventions. In particular, Campbell and Malkus (2011) provide evidence from a three-year randomized control study showing that placing highly trained elementary school mathematics coaches in schools positively affected school-wide student achievement. Notably, this effect was evident only after more than one year of teachers' exposure to coaches. In a recent review of the empirical literature on the effects of teacher induction and mentoring programs, Ingersoll and Strong (2011) also found generally positive impacts of induction and mentoring programs on teacher retention, though they note that there is little evidence of significant impacts on student achievement gains. They also note substantial limitations in the methodologies employed in many of the extant studies in this area.

For example, a pair of studies by Fuller (2003) and Cohen and Fuller (2006) examined the Texas Beginning Educator Support System, a program providing instructional support and mentoring to teachers in their first years of service with the goal of reducing turnover of early career teachers. They found that participants in the program were more likely to remain teachers, even in high-minority and high-poverty schools. However, the study did not account for

differences in teachers who were selected for the program, nor did it control for other factors that may affect teacher retention.

Ingersoll and Smith (2004a, 2004b) used the School and Staffing Survey to examine whether three induction-related measures had an impact on retention. They found that having a mentor in the same field and having regular collaboration with other teachers in the same subject have a positive impact on retention, whereas a reduced teaching schedule or having a teacher aid had a weaker effect. In a subsequent follow-up, they extended the analysis to examine differential impacts by school poverty level and found that teachers in high-poverty schools were less likely to be retained, even though they were equally or more likely to receive support services. These analyses account for teacher and school characteristics that could account for retention, but do not control for unobservable characteristics that may have affected differential selection into the induction programs.

Rockoff (2008) examined whether teachers participating in New York City's 2004 first-year teacher mentoring program were more likely to stay in teaching and have students with improved achievement compared to teachers who are newly hired by the city who have prior teaching experience. He found that having a mentor who previously worked in the same school had a positive impact on retention, and teachers who were given additional hours of mentoring had students with improved mathematics and reading scores. Contradicting evidence from earlier work, he found that matching the mentor and mentee subject areas did not improve teacher or student outcomes.

The largest study of teacher induction, Glazerman et al. (2010), was a randomized evaluation of comprehensive teacher induction for 1,000 teachers in over 400 schools. In this study, first-year teachers in study schools were randomly assigned to receive comprehensive induction for one or two years, and the study examined the impact on teacher practice, retention, and student achievement. Generally the findings were mixed, with no significant impacts on retention, and student achievement impacts only after three years of teaching. As noted by the authors, these findings may be explained in part by the fact that control-group teachers received informal induction (as compared with the formal, comprehensive program offered to treatment teachers) and many treatment group teachers did not participate in all parts of the program.

Other research has explored the impact of programs that provide external professional development or training directly to teachers. A review of the research on professional development interventions of this type by Yoon et al. (2007) indicated that professional development activities can substantially improve teachers' impacts on student achievement. While few of the studies they reviewed met What Works Clearinghouse evidence standards, they

observed substantial average effect sizes across the nine studies that did.[1] Studies that demonstrated effects typically involved a large amount (i.e., around 50 hours) of direct teacher development time. In particular, Yoon and colleagues (2007) note that not one of the nine rigorous studies that they reviewed involved any type of "train-the-trainer" model.

On the other hand, more recent large-scale experimental studies of professional development interventions have failed to show any significant effects of interventions, in spite of the intensive training provided to teachers (Garet et al., 2011; Garet et al., 2008). Moreover, most studies of the impacts of direct professional development to teachers have focused on programs that target teachers' instruction in specific content areas, rather than the type of leadership-focused professional development offered by the LE Fellowship program.

Overall, there are significant gaps in the existing evidence base with respect to understanding the potential impacts of mid-career teacher leadership development programs. Studies related to the impact of mentoring or induction programs have typically focused on novice or early-career teacher mentees. Less work has examined the benefits of developing mid-career teachers as coaches. Moreover, existing studies do not examine the development of active teachers as mentors or teacher leaders, nor do they examine the impact of mentoring programs on the performance of mentors themselves when the mentors are teachers of record. In the area of direct professional development provided to teachers, very little research has examined the impact of development focused on teachers' leadership skills, though other types of professional development have shown some promise of improving teacher performance and student learning. With the expansion of programs such as the LE Fellowship that offer training for mid-career teachers who are themselves actively teaching and who mentor teachers spanning a wide range of prior experience levels, further research is critical to understand whether such programs can achieve their intended goals.

[1] Yoon et al. (2007) report an average effect size of 0.54 on student achievement outcomes. However, many of these effects were measured on local classroom assessments specific to the content or subject matter of the professional development rather than on state-wide annual achievement exams.

3. Description of the Leading Educators Fellowship Program

Overview of the Leading Educators Fellowship Program

The Leading Educators program was founded in 2008 in New Orleans and modeled on the Teaching Leaders program based in London and founded in the same year. In 2011, LE received independent 501(c)(3) status and expanded to Kansas City. The program is based on the theory of change that the racial/ethnic achievement gap can be closed by improving student achievement through supporting teacher leadership activities and by retaining high-quality teachers in high-need schools. The flagship program for the organization is the two-year LE Fellowship program, in which teacher-leaders engage in a number of professional development activities that strengthen their ability to support and provide feedback to their mentee teachers through coaching, professional development, and professional learning communities.

The specific objectives of the program are to (1) increase the leadership skills and capacity of teachers leaders in order to improve student achievement in high-needs schools and (2) retain highly effective teachers in high-need schools by encouraging commitment to the schools and improving the school leadership pipeline. This report provides an evaluation of the extent to which the program met its goals during this period.

Application to the Leading Educators Fellowship Program

The LE Fellowship program is targeted at experienced teachers who work in schools with at least 70 percent of the students eligible for free or reduced-price lunch and who already serve in a leadership role within the school. For example, before they apply, fellows serve as department chairs, grade-level leads, special education coordinators, instructional coaches, or response-to-intervention coordinators. It is not a requirement for fellows to be teachers of record who are directly responsible for teaching students themselves, though some fellows are teachers of record for students.

While the program accepts applications from any teacher who meets the eligibility criteria, it also actively recruits teachers and principals. Recruitment occurs through district and principal partnerships, in-school recruitment, and local recruitment events.

The application process occurs in multiple stages. First, the teachers must show that they meet the eligibility criteria:

1. The teacher must have at least two years of teaching experience.
2. The teacher must work in an open-enrollment school with at least 70 percent of the student population eligible for free or reduced-price lunch.
3. The teacher must have formal or informal responsibility for at least two faculty members in his or her school.

4. The teacher must show evidence of effectiveness from classroom observations or student achievement gains.

If the teacher is eligible, he or she is instructed to complete an online application form and submit a letter of recommendation from their principal. Next, LE staff conducts a 30-minute formal observation of teaching practice. Finally, teachers are invited to an interview and assessment day, where they meet with multiple trained assessors, including LE and local members of the education community. Each stage of the application is scored, with the observation and interviews scored by multiple raters. The final application score is created by combining the individual scores and used to guide decisions about which teacher to accept to the program. While in some years and regions a strict cut-point was used to decide admission, in most instances the total application score was not the only factor considered. In particular, teachers who received low scores on the interview portion of the application process were typically not accepted into the program, even if their total score was relatively high.

Fellowship Training

Once accepted into the program, fellows in training are required to attend a combination of activities: formal professional development sessions, one-on-one leadership coaching, facilitated problem-solving community meetings with other teacher leaders, and site visits to exemplary high-need schools.

The program begins with an induction event, which is followed by a one-week summer intensive training session, where fellows are introduced to the leadership curriculum and learn a majority of the leadership skills. Throughout the school year, fellows meet with leadership coaches to help them integrate the curriculum into their daily practice. Leadership coaches observe the fellows, provide feedback, and aid in developing plans to overcome specific challenges faced by the fellow. Fellows attend problem-solving community meetings, where they discuss solutions to leadership challenges in small teams, and visit schools across the country to observe best practices in person. Finally, fellows complete an action-learning project, where they design and implement an initiative with their mentee teachers to raise student achievement. In the first year of the program, fellows focus on building a foundation of skills for leading teams, facilitating professional development, and coaching. In the second year, the program is focused more on managing change within the school and leading larger initiatives to improve student achievement.

Mentoring Activities

Fellows begin to implement mentoring activities starting in the fall of their first year, because a majority of the leadership content training is included in the induction event and summer intensive session. The mentoring provided by fellows depends on their leadership role in the school and includes coaching, observation of classroom practice with feedback, curriculum co-

planning, data analysis, grade-level meetings, problem-solving communities, modeling instruction, co-teaching, and leading professional development sessions. Fellows meet at least biweekly with their mentees, and at least ten hours per month on mentoring activities.

To summarize, this section described the LE Fellowship program goals, the application process, the types of training received by fellows, and the mentoring activities they lead. In the next section, we discuss the data used for the analysis of the impact of the LE Fellowship program on student achievement gains and teacher retention in Missouri and Louisiana.

4. Data

Our analysis of the impact of the LE Fellowship on student achievement and teacher retention in Missouri and Louisiana used data from LE, including application information and mentee teacher assignments, and combined these data with the Missouri and Louisiana state student and teacher administrative data panel. State administrative data files linked teachers to all of the students they taught in each subject in each school year, allowing us to evaluate the achievement gains of students taught by individual teachers as a function of whether or not they participated in the LE Fellowship program. The LE data files include applicants for three cohorts of fellows in both Missouri and Louisiana, with the first cohort in both states starting the LE Fellowship program in the fall of 2011. Program application files contained the identity of every teacher that applied to the program, the scores for each stage of the application process (online application, observation, interview and assessment day, and written score), and the status of the applicant (e.g., enrolled, graduated, and denied). We also obtained a file that included mentee teacher assignment for each cohort in each year of the program.

We merged the LE files to the state administrative panel data for each state using applicants' names. In Missouri, we had access to state administrative data from 2008–2009 through 2013–2014 school years, whereas in Louisiana we had access to state administrative data from 2009–2010 through 2013–2014. Table 1 provides an overview of the timeline for each cohort in both states by school year. The first cohort applied for the fellowship in the 2010–2011 school year, and teachers who were accepted into the program started their fellowship training in the summer of 2011, completing their first year of the program in the spring of 2012. Fellows from the first cohort are observed for one year after the fellowship program. Fellows in the second cohort are linked to state administrative data after they complete the first and second year of the program. For cohort 3, the data contain only information following the first year of the fellowship program.

Table 1. Leading Educators Fellowship Cohort and Timeline in Missouri and Louisiana

School Year	Cohort 1	Cohort 2	Cohort 3
2008–2009	Pre-applicant*	Pre-applicant *	Pre-applicant *
2009–2010	Pre-applicant	Pre-applicant	Pre-applicant
2010–2011	Applicant	Pre-applicant	Pre-applicant
2011–2012	First-year fellow	Applicant	Pre-applicant
2012–2013	Second-year fellow	First-year fellow	Applicant
2013–2014	Post-fellowship	Second-year fellow	First-year fellow

NOTE: 2008–2009 data available only for Missouri.

We also merged the reported mentee teachers by name to the state administrative panel for these same school years. The primary sample for the analyses of student achievement outcomes was restricted to teachers of students in grades 4 through 8 who had valid state test scores in subject areas that are subject to annual achievement testing in each state.[2] Annual tested subjects in Missouri included mathematics and reading, while in Louisiana they included mathematics, reading, science, and social studies. We also restricted our student achievement analysis sample to include only those districts in the state where the leadership program operated and placed at least one fellow in a school.

The sample used for the retention analysis, which aimed to measure teachers' propensity to remain in teaching, included all fellows and mentees who can be matched to the administrative panel in all districts in their state.[3] In Louisiana, the retention data are restricted to staff who were designated as teachers, so teachers who are promoted to the central office or a leadership position in the school exit the database when they are no longer teachers of record.[4] The Missouri retention data do include promotion information, which allows us to follow teachers as they are promoted to other positions within the district. Also, because many of the teachers participating in the program are located in New Orleans, the Louisiana retention data are influenced by the significant school restructuring that has occurred in the city since Hurricane Katrina.[5]

Table 2 provides sample sizes by school year before and after matching to the state administrative panel on teacher retention and student achievement for program participants (fellows or mentees) in Missouri and Louisiana. "Denied applicants" are those teachers who expressed interest in the program (by initiating the application process) but who failed to advance at one of the three application stages.[6] The sample counts presented in Table 2 combine

---

[2] It is necessary to restrict the analysis of student achievement impacts sample for a number of reasons. We are able to analyze student achievement impacts only for teachers who teach in tested grades and subjects and who have students that can be linked to the teacher in a given year. Missouri tests students starting in grade 3, so all teachers in grades K–2 are excluded from this analysis. We also excluded grade 3 teachers, because the analytic model requires a prior year test score for each student, and students in grade 3 do not have such a score. We examined the high school data in Louisiana and found that only two fellows taught students with achievement scores. We do not have access to high school data in Missouri at this time.

[3] There are differences between the sample used for the analysis of student achievement and the sample used for teacher retention analysis because some fellows do not teach in tested grades and subjects or are not teachers of record who are themselves responsible to teach students.

[4] Leading Educators internal survey data suggest that promotions are fairly common; on Leading Educators' 2014 annual survey, 46 percent of fellows reported receiving a promotion in the previous school year (n=129).

[5] For example, between 2010 and 2014 the Louisiana Recovery School District closed 25 schools, opened 23 new schools, and changed the codes of 21 schools in New Orleans, of the approximately 80 schools it operates (Ferguson, 2014).

[6] As described in the next section, we use denied applicants as a comparison group for fellows in one model specification. Such applicants may share a number of otherwise unmeasured characteristics with fellows, such as the desire to lead teachers and holding positions in schools where the principal supports distributed leadership roles.

information across cohorts to summarize the total number of teachers that are used in the estimation of the program effects.[7,8]

In Missouri, there were between 27 and 56 fellows in each year of the study, compared with between 29 and 37 denied applicants. Fellows list between 86 and 220 mentee teachers across the three years, with more teachers listed in later years (additional mentees may be assigned to participating fellows in each year after entering the program). The program in Louisiana was larger in later school years than in Missouri, admitting 84 fellows in the 2013–2014 school year. In accordance with having more fellows, there were also more mentees in Louisiana in later years. There was no information available on intended mentee teachers for denied applicants in either state.

Table 2. Sample Counts

	Missouri		Louisiana	
	Program Participants	Denied Applicants	Program Participants	Denied Applicants
School Year 2011–2012				
Enrolled fellows	27	37	18	47
Enrolled fellows matched to retention data	16	31	16	38
Enrolled fellows matched to achievement data	6	8	11	18
Enrolled mentees	86	N/A	51	N/A
Enrolled mentees matched to retention data	67	N/A	42	N/A
Enrolled mentees matched to achievement data	55	N/A	25	N/A
School Year 2012–2013				
Enrolled fellows	35	36	40	64
Enrolled fellows matched to retention data	19	17	31	47
Enrolled fellows matched to achievement data	7	12	12	21
Enrolled mentees	99	N/A	151	N/A
Enrolled mentees matched to retention data	83	N/A	121	N/A
Enrolled mentees matched to achievement data	77	N/A	50	N/A
School Year 2013–2014				
Enrolled fellows	56	29	79	95
Enrolled fellows matched to achievement data	16	14	23	21
Enrolled mentees	220	N/A	309	N/A
Enrolled mentees matched to achievement data	120	N/A	111	N/A

The sample size of fellows and mentees dropped significantly once the data were merged to the state administrative panel on teacher retention and student achievement in both states. There was a decrease in the number of participating teachers who could be identified in the teacher-level retention file, with 16 first-year fellows in 2011 and 31 first-year fellows in 2012 in Louisiana, and 16 first-year fellows in 2011 and 19 first-year fellows in 2012 in Missouri.[9]

[7] For example, the 2012–2013 sample counts of the number of enrolled fellows combines second-year fellows in cohort 1 with first-year fellows in cohort 2.

[8] There are no sample sizes displayed in the 2013–2014 school year for teacher retention because 2013–2014 is the last year of data available for analysis, and we are not able to observe retention in the following school year.

[9] We excluded new teachers in 2013–2014 from the retention analysis file because we do not have data from the following school year to calculate whether those teachers were retained.

Similarly, across each of the three cohort-years and across both states, anywhere from six to 23 fellows could be matched to students in grades 4–8 in the administrative data.[10] There was a drop in the sample of mentee teachers as well, though many more of them were found in the state database teaching students in tested grades and subjects.[11]

Tables 3 and 4 summarize the characteristics available in the state administrative panel for teachers and for the students they taught, respectively, for different groups of teachers in Missouri.[12] Teacher characteristics from the state panel include teachers' race, gender, highest degree earned (e.g., bachelor's degree, master's degree, doctorate), and indicators for years of experience. Student-level information includes demographic characteristics, such as race, gender, free or reduced-price lunch eligibility, special education status, limited English proficiency (LEP) status, and building mobility status. Student test scores were available for mathematics and reading, and they are standardized by subject, grade, and year.

The different populations of teachers detailed in Tables 3 and 4 include fellows, denied applicants, mentee teachers and district teachers. "Mentee teachers" are those who were selected by fellows to mentor in a given school year, whereas "district teachers" includes all teachers who worked in school districts where the LE Fellowship program operated. The tables also show the difference in average characteristics and statistical significance of this difference between fellows versus denied applicants, and mentees versus district teachers. The sample of teachers presented in the tables is restricted to fellows and mentees who taught students in tested subjects and who could be linked to student achievement scores.

Overall, fellows were 73 percent female, and one-third of fellows had a master's degree. Compared denied applicants, fellows were less likely to be female, more likely to be Hispanic, and tended to have fewer years of experience. Mentee teachers were less likely to be female, more likely to be black, more likely to be Hispanic, and had fewer years of experience than district teachers.[13]

Comparing the student characteristics of subgroups of teachers presented in Table 4, fellows taught students with higher mathematics and reading test scores than denied applicants or mentee teachers, but lower test scores than other teachers in the district. Fellows also taught students

---

[10] In each year, about one-third of the fellows listed were excluded because they taught grades K–3 or high school students. The remaining teachers who were not included in the analysis were not teachers of record. A handful of fellows' names could not be found in the state administrative database, even after checking alternative spellings.

[11] The poor matching rates for fellows is due to program design, which does not require teachers applying to the program to be teachers of record, instead requiring that they hold leadership positions in the school. On the other hand, because the goal of the program is to improve student achievement, mentee teachers were selected based on whether they were teachers of record.

[12] The characteristics are summarized for the sample used in the analysis of student achievement.

[13] Because these statistics are based on the sample used in the analysis of student achievement, they reflect only the demographic composition of teachers who were matched to student achievement data, and may not represent program participants as a whole.

who were less likely to be black, more likely to be Hispanic, and more likely to be LEP than students of denied applicants.

Table 3. Teacher Demographic Characteristics, by Teacher Subgroups, Missouri

	Fellows	Denied Applicant	Difference Fellows and Denied Applicants	Mentee Teacher	District Teachers	Difference Mentees and District Teachers
Total salary	49,017	47,262	1,755**	44,688	48,527	–3,838**
% female	73.0	85.0	–12.0**	79.5	83.0	–3.5**
% Asian	1.3	0.0	1.3**	1.4	0.9	0.5**
% black	34.1	31.3	2.9	32.7	12.3	20.4**
% American-Indian	0.0	0.0	0.0	0.3	0.2	0.1
% Hispanic	23.6	10.5	13.1**	2.0	1.4	0.6**
% Pacific Islander	0.0	0.0	0.0	0.0	0.0	0.0**
% multi-race	0.0	0.0	0.0	0.3	0.1	0.2**
% bachelor's	66.8	66.8	0.0	67.5	39.4	28.1**
% master's	33.2	32.8	0.4	29.1	58.2	–29.0**
% doctorate	0.0	0.0	0.0	1.2	0.3	0.9**
% education specialist	0.0	0.4	–0.4**	1.4	1.7	–0.3**
% other degree	0.0	0.0	0.0	0.3	0.3	0.1
% experience 1 year	3.8	2.0	1.8**	11.8	8.1	3.7**
% experience 2 years	3.1	8.7	–5.6**	12.6	7.0	5.6**
% experience 3 years	14.0	5.0	9.0**	11.0	6.2	4.8**
% experience 4 years	12.3	12.6	–0.3	6.3	5.0	1.3**
% experience 5 years	10.9	7.0	3.9**	6.2	5.6	0.6**
% experience 6 years	7.9	0.1	7.7**	5.1	4.9	0.2
% experience 7 years	4.7	0.2	4.5**	2.4	4.9	–2.5**
% experience 8 years	9.5	10.1	–0.6	5.0	4.7	0.3
% experience 9 years	0.0	6.0	–6.0**	3.9	4.2	–0.3**
% experience 10 years	0.0	5.4	–5.4**	5.1	3.7	1.3**
% experience 10+ years	33.9	43.0	–9.0**	30.7	45.7	–15.0**
N of mathematics teachers	15	20		116	4,048	
N of reading teachers	15	17		126	4,120	

NOTE: Demographic characteristics for teachers used in the analysis of student achievement, mathematics, and reading teachers combined. Summary statistics combine information from the 2011–2012 and 2013–2014 school years. Statistical significance calculations presented for fellows compared with denied applicants, and for mentees compared with district teachers. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

Table 4. Student Demographic Characteristics, by Teacher Subgroups, Missouri

	Fellows	Denied Applicant	Difference Fellows and Denied Applicants	Mentee Teacher	District Teacher	Difference Mentees and District Teachers
Standardized mathematics score	–0.175	–0.423	0.248**	–0.580	–0.129	–0.451**
Standardized reading score	–0.300	–0.541	0.267**	–0.481	–0.121	–0.360**
% free- or reduce-price-lunch-eligible	90.4	84.8	5.6**	85.9	54.7	31.2**
% American-Indian	0.4	0.1	0.3	0.2	0.4	–0.1**
% Asian/Pacific Islander	3.4	0.6	2.8**	1.7	2.1	–0.4**
% black	49.6	54.3	–4.6*	70.0	32.8	37.1**
% Hispanic	39.1	32.5	6.6**	16.9	11.2	5.7**
% multi-racial	0.1	0.1	0.0	1.1	1.6	–0.5**
% female	54.1	49.4	4.7*	50.3	49.4	0.8
% special education	6.2	11.8	–5.6**	8.8	10.2	–1.5**
% ltd. English proficiency	36.9	23.7	13.1**	13.9	6.3	7.6**
% in building for <1 year	4.9	7.0	–2.1*	6.5	5.3	1.2**
% Missing lagged mathematics score	0.0	0.1	–0.1	0.2	0.1	0.1
% Missing lagged reading score	1.9**	0.3	1.7**	0.5	0.2	0.2**
N	675	959		7,602	215,073	

NOTE: Demographic characteristics for student used in the analysis of student achievement. Summary statistics are averages across all student-year observations for each group of teachers in both subjects. Statistical significance calculations presented for student of fellows compared with students of denied applicants, and for students of mentees compared with students of district teachers. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

Tables 5 and 6 display a similar set of teacher and student characteristics, this time based on the analysis sample in Louisiana. There are a few differences between the two states in terms of the information collected in the administrative panel. In particular, Louisiana does not allow teachers to have less than a bachelor's degree, and includes an additional education category called "master's plus 30," which gives credit to teachers who have taken graduate course work after obtaining a master's degree. In addition, Louisiana was unable to provide information on the teacher's years of experience in each school district. Therefore, we calculated two approximate measures of experience: (1) an indicator for the teacher appearing for the first time in the school dataset in a particular year and (2) a proxy for teacher experience that is defined as the teacher's salary relative to the average salary in the school by education category.[14] The primary difference in the student characteristics across the two administrative panels is that Louisiana fields annual tests in four subjects: mathematics, reading, science, and social studies. In addition, Louisiana collects more detailed information about students, including the age of the student (which we convert to relative age of the student in the classroom), whether the student is retained in the grade, indicator for whether the student is classified as gifted, the total number of disciplinary incidents, and the total number of days suspended (in or out of school).

[14] Since teacher pay scales at this time in Louisiana were based on education level and years of experience, this measure increases according to the teacher's tenure.

Table 5. Teacher Demographic Characteristics, by Teacher Subgroups, Louisiana

	Fellows	Denied Applicant	Difference Fellows and Denied Applicants	Mentee Teacher	District Teachers	Difference Mentees and District Teachers
% female	93.9	90.9	3.0	82.9	89.2	–6.3**
% Asian	0.0	0.0	0.0	1.7	2.1	–0.3
% black	12.2	60.6	–48.4**	34.6	31.1	3.5
% American-Indian	0.0	0.0	0.0	0.0	0.1	–0.1**
% Hispanic	0.0	0.0	0.0	0.7	1.6	–0.9
% Pacific Islander	0.0	0.0	0.0	0.0	0.0	0.0**
% multi-race	2.0	0.0	2.0	0.0	0.2	–0.2**
% master's	24.5	28.8	–4.3	16.1	23.6	–7.6**
% master's + 30	0.0	7.6	–7.6**	0.3	4.6	–4.3**
% doctorate	0.0	0.0	0.0	2.1	0.4	1.7**
% education specialist	0.0	0.0	0.0	0.0	0.6	–0.6**
% first-year indicator	2.6	1.9	0.7	11.0	10.9	0.2
Continuous experience proxy	–441	1,190	–1,631	–428	–16	–411**
N of mathematics teachers	7	18		53	4,487	
N of reading teachers	23	31		73	6.878	
N of science teachers	9	12		38	3,723	
N of social studies teachers	5	13		45	4,218	

NOTE: Demographic characteristics for teachers used in the analysis of student achievement, mathematics, reading, science, and social studies teachers combined. Summary statistics combine information from the 2011–2012 and 2013–2014 school years. First-year indicator refers to first year in the 2009–2010 to 2013–2014 administrative panel. Continuous experience proxy is salary of the teacher relative to the school average, separate by experience category. Statistical significance calculations presented for fellows compared with denied applicants, and for mentees compared with district teachers. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

Table 6. Student Demographic Characteristics, by Teacher Subgroups, Louisiana

	Fellows	Denied Applicant	Difference Fellows and Denied Applicants	Mentee Teacher	District with Applicants	Difference Mentees and District Teachers
Standardized mathematics score	–0.030	–0.333	0.303**	–0.316	–0.008	0.308**
Standardized reading score	–0.235	–0.227	0.008	–0.375	–0.062	0.313**
Standardized science score	–0.412	–0.357	0.054	–0.307	–0.076	0.231**
Standardized social studies score	–0.138	–0.242	0.242	–0.30	–0.04	0.264**
% free- or reduce-price-lunch-eligible	93.8	88.6	5.2**	93.3	76.2	17.1**
% American-Indian	0.1	0.3	–0.2	0.2	0.3	–0.1**
% Asian/Pacific Islander	0.4	2.4	–1.9**	0.7	2.4	–1.7**
% black	91.2	83.5	7.8**	86.4	58.1	28.3**
% Hispanic	2.6	6.2	–3.5**	4.0	6.8	–2.8**
% multi-racial	0.4	0.7	–0.3	0.7	1.0	–0.2**
% female	48.1	48.4	–0.3	47.7	48.7	–1.0**
% special education	17.1	14.1	3.0**	15.3	11.0	4.2**
% limited English proficiency	1.0	3.1	–2.1**	1.4	3.6	–2.2**
% in building for <1 year	95.4	96.7	–1.3**	94.0	95.9	–1.9**
Relative age	0	0	0.0	0	0	0.0
% gifted	3.4	5.4	–2.0**	4.0	7.4	–3.4**
% retained	3.2	4.1	–0.9**	4.8	3.7	1.1**
N of disciplinary incidents	0.7	0.3	–0.4**	0.6	0.5	–0.1**
N of suspension days	1.4	0.7	0.7*	1.4	1.4	0.0
N of students, math	249	820		3,060	242,454	
N of students, reading	1,220	1,667		3,960	324,573	
N of students, science	728	508		2,365	240,749	
N of students, social studies	159	527		2,942	245,213	

NOTE: Demographic characteristics for student used in the analysis of student achievement. Summary statistics are averages across all student-year observations for each group of teachers in both subjects. Statistical significance calculations presented for student of fellows compared with students of denied applicants, and for students of mentees compared with students of district teachers. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

As was the case in Missouri, fellows in Louisiana were predominantly women with a bachelor's degree. Fellows in Louisiana were less likely to be black than denied applicants. Mentee teachers were less likely to be female and more likely to have a master's degree than other teachers in the district.[15]

When comparing the student characteristics of Louisiana teachers, we see that fellows taught students with higher mathematics and reading test scores than denied applicants or mentee teachers, but lower scores than students of other teachers in the district. Fellows also taught students who were more likely to be black, more likely to receive free or reduced-price lunch, less likely to be gifted, and who had more disciplinary infractions compared to students of denied applicants.

[15] Because these statistics are based on the sample used in the analysis of student achievement, they reflect only the demographic composition of teachers who were matched to student achievement data, and may not represent program participants as a whole.

Overview of the Leadership Competency Rubric

The purpose of the training received by fellows was to master a core set of leadership knowledge and skills that could be used to improve student outcomes. To measure improvement in these skills and to create a roadmap for teacher leadership development, LE consulted with experts and reviewed the research literature in order to develop the Leadership Competency Rubric (LCR). The LCR assesses leadership skills and was used to evaluate fellows throughout the course of the two-year fellowship program. Here we describe the LCR, and later we present a descriptive analysis of how the LCR scores changed over the course of the LE Fellowship program for various subgroups.

The version of the rubric used in the 2011–2012 to the 2013–2014 school years contained four strands: (1) Core Beliefs and Mindsets, (2) Management of Self and Others, (3) Cultural Leadership, and (4) Instructional Leadership. Each strand contained eight competencies, for a total of 32 ratings. Fellows were evaluated using the LCR three times during the LE Fellowship program: round 1 at the beginning of the program, round 2 at the end of the first year, and round 3 at the end of the second year. The evaluations were completed by the fellow, the principal, and up to four colleagues of the fellow during each round. The LCR is scored on a six-point rating system, with the possible scores including No Evidence, Explore, Embark, Extend, Embed, and Empower. By the end of the two-year program, fellows were expected to show improvement in all competencies.

Descriptive Statistics for Rating Score Changes

Figure 1 displays the overall Leadership Competency Rubric scores for all fellows, by type of rater for each round. The top left cell shows the self-reported ratings, the top right cell shows the principal ratings, the bottom left cell shows the ratings by colleagues, and the bottom right cell shows the average of self-reported ratings and principal ratings. All groups showed improvement in ratings across rounds, and for all groups the difference in ratings between round 1 and round 2, and round 2 and round 3, are statistically significant at the 5 percent level.[16] Because colleague ratings were available only in rounds 2 and 3, for the remainder of the descriptive analysis we focus on averages of the self-reported and principal ratings.

[16] While the statistics presented here do not account for the changing composition of fellows (due to some fellows leaving the program and new fellows joining), we observe the same trend in scores when we restrict the samples to only those fellows who are present in all three rounds.

Figure 1. Leadership Competency Rubric Scores, by Rater and Round

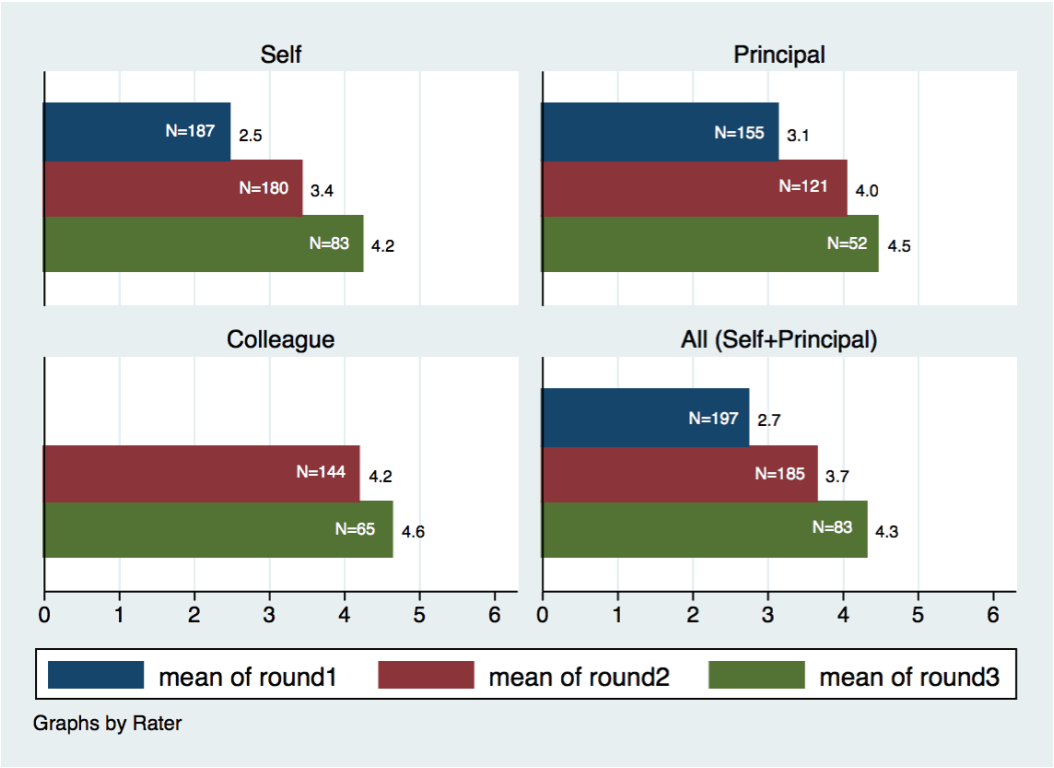
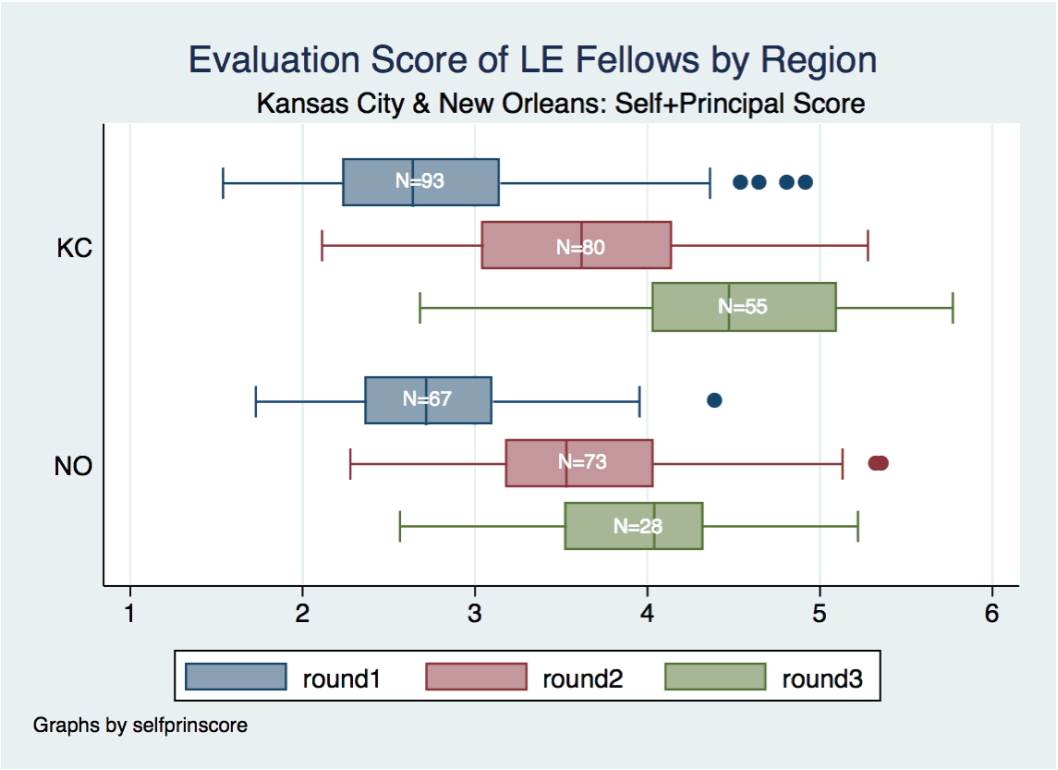


Figure 2 shows the ratings scores of fellows by region across the three rounds. The boxplot displays the distribution of ratings, with median in the middle of the box, and the 25th and 75th percentiles of the ratings at the edges of the box. The "whiskers" indicate extreme low and high values, and the dots indicate outlier values. From this figure it is evident that overall rubric scores improved in both Missouri and Louisiana, and that there was a significant amount of variation in both the initial and final rubric scores, especially in Missouri.[17]

[17] The decrease in the sample size across rounds is partially due to fellows leaving the program, but in large part it reflects the fact that there are a larger number of fellows in later program cohorts, and with the rounds being spaced one year apart, later cohorts have only been scored in the earlier rounds.

Figure 2. Leadership Competency Rubric Scores, by Region and Round



Finally, Figures 3 and 4 show the Leadership Competency Rubric scores by round and year of entry into the LE Fellowship program for Missouri and Louisiana, respectively. Focusing first on Missouri, we observe that the scores improved across rounds for each cohort of fellows, with somewhat larger improvements in the 2011 fellows compared with the 2012 fellows, with 2013 fellows showing similar early improvement trends as the 2012 fellows. In Louisiana, the LE Fellowship program did not collect round 1 data, but comparing across the cohort years for rounds 2 and 3, the 2012 cohort realized larger gains on the rubric than the 2011 cohort. The 2013 Louisiana cohort also showed significant gains between the first two rounds.[18]

[18] We also examined the LCR scores by sociodemographic characteristics and found that scores improved at a similar rate for male and female fellows and for white and nonwhite fellows. In addition, we found similar improvement rates when breaking out the overall score into the four competencies.

Figure 3. Leadership Competency Rubric Scores, by Year and Round, Missouri

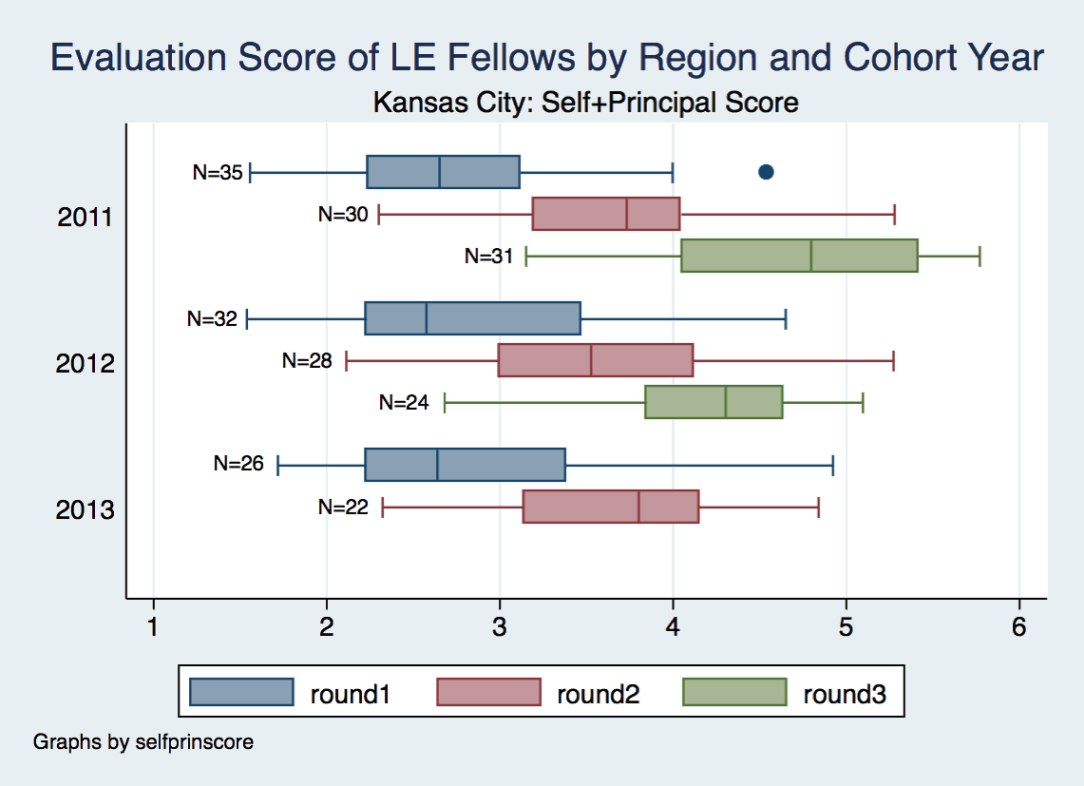
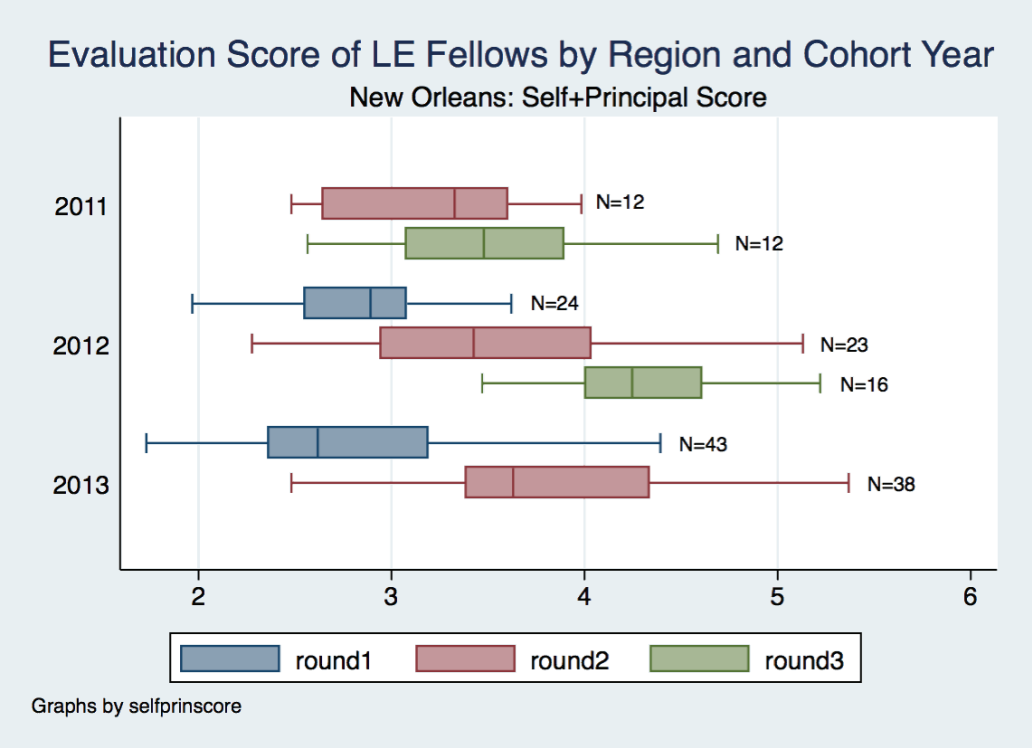


Figure 4. Leadership Competency Rubric Scores, by Year and Round, Louisiana



Overall, we observe that fellows and mentee teachers differed in a number of ways from teachers in the districts that may have affected student achievement gains. In addition, although fellows and denied applicants may have been similar on unmeasured characteristics, such as a desire to lead teachers and teaching in schools where the principal supports distributed leadership roles, the result in Tables 2–5 demonstrate that they differed on a number of key characteristics that may relate to their students' achievement gains. We also found that LCR scores improved for fellows across all regions, in all years, providing suggestive evidence that fellows are improving their leadership skills and that the LE Fellowship program shows promise for achieving its goals of impacting student achievement and teacher retention. In the next section, we describe our analytic methods, including how we controlled for differences in teacher, student, and school-wide characteristics when analyzing the impact of the program on student achievement.

5. Analytic Methods

In this section, we discuss our empirical approach to evaluating the impacts of the LE Fellowship program. We considered evidence of the program's impact on both student achievement gains and teachers' retention, and evaluated possible program effects both for fellows and for the teachers they mentor.[19] Our evaluation of program effects on student achievement gains used a range of rigorous methods, including a comparison of fellows and applicants who were not accepted into the program (with controls for application scores), a propensity score matching (PSM) method that compares fellows and mentees with teachers who appear equally likely to have participated in the program, and a matching method that compares the performance of fellows and mentees relative to peers who had similar instructional impacts in a prior year. Each method that we employed has both strengths and limitations, and collectively they present a more robust picture of the program's possible effects on student achievement gains.

All specifications measure the impact of participating in the LE Fellowship program for one or more years, calculating the average impact across multiple years. For fellows, we combine the effect of first-year fellows, second-year fellows, and those who graduated from the program.[20] For mentees we combine the effects of being a mentee for one or multiple years, and we consider mentees "treated" if they were ever a mentee. While this approach is necessary due to the small number of program participants we can match to the student administrative panel and has its limitations, we believe it is justified given that student achievement scores are measured at the end of the school year. By that point, even first-year fellows will have participated in the program activities throughout the year and have already started to implement their skills with their mentee teachers.

Student Achievement Analysis

To evaluate the impacts of the LE Fellowship program on student achievement gains, we considered three complementary modeling approaches:

1. a comparison of fellows to denied applicants with controls for application rating scores[21]

[19] In preliminary work, we also explored possible school-wide spillover effects of fellows, distinct from any effects observed among students of fellows or mentees. However, we did not identify statistically significant school-wide spillover effects.

[20] Thus we assume that the impact of fellows on student achievement is equal for current fellows and graduates. In future reports when additional data are available, we will consider more flexible specifications that allow the impact to vary by LE Fellowship program graduation status and mentee exposure to graduated fellows.

[21] In preliminary work, we considered employing a regression discontinuity approach, given the existence of the score cutoffs. Regression discontinuity is a widely recognized method and rigorous method for program evaluation.

2. a PSM approach
3. an approach that matches teachers solely according to their pretreatment effectiveness at raising student achievement.

Each modeling approach compared the student achievement gains of "treated" fellows or mentee teachers who participated in the LE Fellowship program with those of a "control" comparison group of teachers who did not, and each approach has distinct advantages and limitations. Because each modeling approach may be subject to different biases, using all three models can provide a test of the robustness of our findings.[22]

In Table 7, we describe the three modeling approaches and the advantages and potential limitations of each. Across all three modeling approaches, when comparing the student achievement gains of treated teachers (either fellows or mentees) and comparison teachers, we control for a wide range of observable student, school, and teacher characteristics that may have impacted student achievement gains. In the following subsections, we discuss each modeling approach in greater detail.

Table 7. Methods, Advantages, and Potential Limitations of Modeling Approaches

Modeling Approach	Method	Advantages	Limitations
Comparison of fellows to denied applicants	Compared the instructional effectiveness of fellows and program applicants who were not accepted, with controls for screening rating scores.	If ratings capture all fellow selection effects, including their effectiveness and likelihood of improvement, then this method approximates random assignment of fellows to treatment.	Ratings may not reflect all selection criteria used to screen applicants; rejected applicants may differ from fellows in unobserved ways.
Propensity score matching	Compared fellows or mentees to other teachers who appear similarly likely to have been treated, based on observable characteristics.	Allows for investigation of effects for both fellows and mentees (who lack an applicant comparison group).	Matching on (and controls for) observable characteristics may not fully account for unobserved program selection effects.
Matching on pretreatment effectiveness	Compared fellows or mentees to peers with similar impacts on student learning in the year prior to treatment.	Ensures that treatment and control teachers are equivalent prior to treatment in terms of our outcome of interest.	The need for baseline data limits our sample to teachers with prior-year data on student achievement gains.

Unfortunately, because the sample size of applicants in any one year is very small, the regression discontinuity approach is powered to detect only unreasonably large effect sizes.

[22] Due to the variety of channels by which teachers and principals are recruited into the LE Fellowship program, it is difficult to ensure that any comparison group we identify is truly equivalent to the treated teacher fellows, or, to a lesser extent, their mentees.

Comparison of Fellows and Denied Applicants

In this specification, we compared the performance of students of fellows with those of applicants who did not enter the program, by predicting student achievement in each year as a function of students' prior-year achievement, student characteristics, school composition characteristics, teacher characteristics, year effects, grade effects, an indicator that is set to 1 in all years after a teacher *applied* to the LE Fellowship program, and an additional indicator (i.e., interaction term) that is set to 1 in all years after a teacher enters the program as a fellow. With this specification, we estimated the differential achievement gains of students of fellows, over and above that of students of applicants. To control for possible differences in the achievement gains of students of accepted and rejected applicants that are captured by the screening process, we also included in the regression the application score used to screen and determine admission into the LE Fellowship program, and an indicator for whether the applicant received a low score on any single screening interview. The Technical Appendix includes the specification of this regression equation.

A key assumption underlying the comparison of fellows and denied applicants is that the two samples of teachers are comparable in terms of their effectiveness and their improvement trajectory, when controlling for selection ratings. This is likely to be the case only to the extent that the application score that determined whether they were accepted into the program accurately reflects all of the selection criteria that went into determining which applicants were accepted. While rating cutoff scores did exist in both Missouri and Louisiana in most years of the program, these were not strictly enforced, with some applicants close to the cutoff being accepted and others denied. In addition, having a low score on any single screening interview appears to have been a factor in the rejection of some applications, and as such we have included a control for that information as well in this modeling approach. Overall, even if our controls for application score capture most of the difference in the performance and improvement trajectories of fellows and denied applicants, there may still be important unobservable differences between fellows and denied applicants, particularly for those with rating scores close to the cutoff. Nevertheless, this modeling approach is uniquely helpful in that it compares two groups of teachers with a similar expressed interest in developing their teacher leadership practices.

Because we do not have corresponding information about potential "applicant" mentees in the way that we do for fellows, we were unable to use this modeling approach to evaluate program effects for mentees.

Propensity Score Matching

Our goal in all of our analyses was to identify an appropriate untreated comparison group of teachers that could serve as a counterfactual for how LE Fellowship program participants would have performed had they not experienced the program. The second modeling approach that we employed was a PSM technique, which allowed us to address this challenge by identifying

comparison teachers that were equivalent to fellows (or mentees) in terms of their observed likelihood of participating in the program (Dehejia and Wahba, 2002). We conducted separate PSM analyses for fellows and mentee teachers. A key benefit of this approach was that we could also use it to estimate the impact of the program on mentees. We were unable to examine mentee impacts with the approach described previously because the program did not collect information about potential mentees of denied applicants.[23]

In the PSM approach, we first identified a matched comparison sample that appeared similarly likely to have participated in treatment (i.e., to have been a fellow or a mentee), and then compared outcomes between "treated" teachers and the matched "control" teachers. PSM techniques can improve our ability to make claims of causal inference to the extent that the observable characteristics of teachers that we can measure in our data and that are used in the matching process represent the complete set of factors that made teachers likely to become a fellow or mentee. In this case, we are unlikely to have had data that capture all of the information that may have led teachers or schools to participate in the LE Fellowship program.[24] Nevertheless, matching likely improved our ability to generate an accurate estimate of LE Fellowship program effects by increasing the comparability of our treated and comparison groups.

A variety of matching algorithms may be used in the identification of an untreated comparison group (Caliendo and Kopeinig, 2008). In this case, we found that many-to-one "nearest-neighbor" matching with replacement yielded the most consistently balanced matches for fellows and mentees. In particular, we matched each treated LE teacher in each year with up to ten comparison teacher-years that had similar observed propensity to receive treatment. This limited our selection of matches to include only those whose propensity for treatment was very close to that of treated observations. Any control observations that served as matches for multiple treatment observations were weighted based on the number of times they were selected as matches.

Our model for estimating teachers' propensity to receive treatment included teachers' grade levels, years of prior experience, race, gender, and degree levels; the demographic and prior achievement characteristics of the students they were assigned to teach; school characteristics, including the percentage of students in the school eligible for free or reduced-priced lunch and the average school-level test score in the prior year; classroom characteristics, such as the percentage in the classroom eligible for free or reduced-priced lunch and year fixed effects.

---

[23] In preliminary analysis, we considered using teachers in the same grade and subject as the denied applicant as a proxy for their potential mentees, but we found using this method for fellows to be a poor approximation of the types of teachers who are chosen as mentees in this program.

[24] For example, we do not have information on the process by which schools were identified by LE as possible collaborators, or how mentee teachers were selected by fellows or school officials.

Using this model, we identified teacher-year observations as matches for treated LE fellow or mentee teachers in each year.

While fellows and mentees were quite different from most other teachers in our overall sample, we found that, across all subject areas, there was sufficient overlap in teachers' propensity to receive treatment to identify a matched sample of teachers that were comparable to treated teachers. We provide greater detail on this methodology in the Technical Appendix, and we include in Appendix A descriptive tables showing the comparability of treatment and matched samples across a range of observable characteristics that were used in matching.

Matching on Pretreatment Effectiveness

The previous methods for assessing the impacts of the LE Fellowship program adjust in different ways for observable differences in teacher characteristics and school contexts. However, it may still be the case that treatment and control teachers differ in terms of their effects on student achievement gains in ways that not fully accounted for. As an alternative approach, we attempted to ensure the similarity of treatment and matched control teachers by matching them according to their measured effects on their students' achievement gains in the school year prior to treatment. In other words, we identified control teachers whose measured impacts on student learning were similar to those of fellow or mentee teachers in the year prior to program participation. Control teachers were selected from a sample of teachers who were present in both the pretreatment year and in the first year in which each fellow (or mentee) participated in the program.

Establishing "baseline equivalence" in this way corresponds to federal What Works Clearinghouse standards for quasi-experimental research[25] by ensuring that treated teachers are matched with a sample of control teachers who were equivalent in terms of our outcome of interest in the year prior to receiving treatment. A limitation of this approach, however, is that it excluded teachers for whom data on student achievement gains were not available in the year prior to treatment. This restricted our available sample, excluding in particular mentee teachers who were in their first year of teaching in the district when they first received mentoring. Another potential limitation is that the measure we use to estimate the teacher's effect on students achievement gains is relatively unstable, with typical year-to-year correlations of 0.2–0.7 (McCaffrey et al., 2009). Teachers are matched based on measured instructional effectiveness at baseline, but because this measure is imprecise, it is possible that matched and comparison teachers' true instructional effectiveness at baseline varies in ways that we do not observe. We provide a more detailed description of this modeling approach in the Technical Appendix.

[25] Specifically, our approach represents an "adjacent cohort" design measuring treatment effects for teacher-level "clusters" that serve different groups of students in each school year.

Power Analysis

Given our relatively small sample of fellows in both states, our analyses were in several cases powered only to detect large program impacts on their students' achievement gains. To illustrate this, we show in Appendix Table A6 the types of minimum detectable effect sizes (MDEs) that we would be reasonably powered to detect across a range of sample sizes that are present in our analyses. Overall, fellows were not well powered to detect effect sizes in a range that we would consider likely, with MDEs ranging from 0.174 to 0.356 across our typical fellow-year sample sizes (i.e., 5 to 20), given a goal of being powered to detect effects 80 percent of the time. MDEs remain large even if our goal is to be powered to detect fellows effects only 50 percent of the time, with a corresponding range from 0.115 to 0.236. This means that it is very possible that we fail to detect any true difference in the learning gains of fellows, if those true differences are not large. For mentees, however, for whom teacher-year sample sizes ranged from around 40 to around 200, we do in some cases have sufficient power to detect more modest program effects, ranging from 0.123 to 0.055, respectively, at 80 percent power. Thus, our power analyses indicate that we were reasonably likely to adequately assess the potential for moderate effect sizes among mentees (particularly in Missouri), but unlikely to adequately assess the potential for moderate effect sizes among fellows.

Teacher Retention Analysis

One of the goals of the leadership program was to improve teacher retention rates, and in particular to encourage teachers to remain in high-poverty schools. We first examined whether there was any evidence that the program was associated with differential retention rates using descriptive analyses in both Missouri and Louisiana We report one- and two-year retention rates for fellows, denied applicant mentees, and all teachers in the district. For each of these groups of teachers, we report three different retention statistics: (1) same school retention, (2) retention in teaching, and (3) retention in high-poverty schools. When examining retention rates in high-poverty schools, we limited our sample to those teachers who were initially working in a high-poverty school.[26]

To more rigorously evaluate the potential retention impacts of participation in the LE Fellowship program, we implemented a PSM methodology similar to that described for the evaluation of student achievement impacts.[27] Specifically, for each treated teacher in their initial year of treatment, we identified a set of up to ten nearest neighbor matches that appeared equally likely to have been treated in that period. We then estimated models that predict the probability

[26] For Missouri, we were also able to access information about promotions to leadership positions, such as rates of promotion to principal or central office roles.

[27] Because of the small number of fellows in our sample, we were able to implement the regression analysis only for mentees.

that treatment and control teachers were retained in the next one or two school years, while controlling for observable student, teacher, and school composition characteristics, as well as year effects.

As we did for the modeling approaches related to student achievement outcomes, we provide a more detailed description of our retention modeling approach in the Technical Appendix. In the following section, we describe the results of our analyses of program impacts.

6. Results

In this section, we detail the results of our analyses of LE Fellowship program impacts on participating teachers students and on the career paths of teachers themselves. Overall, we observed mixed results with respect to the program's impacts on student achievement gains thus far. In some subject areas and model specifications, fellows appeared more effective than comparison teachers, while in others, fellows appeared to perform worse. While we observed no clear evidence of positive or negative effects on mentees in Missouri, where we had the largest sample of potentially impacted teachers, we did find that mentees in Louisiana appear to have performed better in some subjects as a result of the program.

With respect to the preliminary analysis on teacher retention, we found mixed results. Mentees in Missouri were less likely to stay in the same school or in teaching than comparison teachers, whereas mentees in Louisiana were more likely to stay in teaching, but less likely to stay in the same school. Below, we present the findings from these analyses.

Student Achievement Impacts

First, in Table 8, we summarize the results of the regression equations measuring program impacts on students' mathematics achievement gains in Missouri. In the first column, we show coefficients from the specification that compares fellows with other program applicants. In this case, the estimated differential impact of fellows on student achievement is positive relative to that of denied applicants (when controlling for application ratings) but not statistically significant, with an effect magnitude of approximately 0.036 standard deviations in student test scores.

Other point estimates related to mathematics achievement among students of fellows in Missouri from Table 8 reveal mixed results. Column 2 indicates that we observed no significant effect of fellows' program participation in the basic PSM specification. However, as shown in Column 3, a model that matches fellows to peers with comparable pretreatment instructional impacts indicated a significant negative effect of program participation, with a large negative coefficient of –0.276. While the sample of fellows matched on pretreatment performance in Column 3 is somewhat smaller than that of Columns 2 or 1, this difference in estimated effects across model specifications indicates considerable inconsistency across model specifications. Columns 4 and 5 show results for models estimating program impacts on mentee math teachers in Missouri who are matched on observables or on pretreatment performance, respectively. In both cases, we observe no significant effect of the program on mentee performance.

Table 8. Impact of the Leading Educators Fellowship Program on Mathematics Achievement in Missouri

	Fellows vs. Applicants	Fellow PSM	Fellow Prior Effects PSM	Mentee PSM	Mentee Prior Effects PSM
Fellow	0.036	–0.019	–0.276**		
	(0.068)	(0.047)	(0.071)		
Applicant	0.101				
	(0.081)				
Mentee teacher				0.005	–0.038
				(0.027)	(0.041)
Application score	–0.000				
	(0.006)				
Student demographics	Y	Y	Y	Y	Y
School year and grade indicators	Y	Y	Y	Y	Y
Classroom composition	Y	Y	Y	Y	Y
School composition	Y	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y	Y
Observations	210,126	4,305	7,134	23,590	24,577
R-squared	0.67	0.59	0.61	0.65	0.70
N of fellow-years (F)	23	24	18		
N of denied applicant–years (DA)	23				
N of mentee teacher–years (MT)				159	96

NOTE: Standard errors in parentheses, clustered at the teacher-by-year level. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

In Table 9, we summarize corresponding results related to program impacts on student reading achievement gains in Missouri. Here, we observe marginally significant positive impact estimates for fellows in both the model specification that compares their performance to that of denied applicants, and in a model that compares their performance to that of a matched sample of peers. However, as was the case in mathematics in Missouri, the estimated impacts in a specification that matches fellows according to pretreatment effectiveness is inconsistent with the other model results, with a nonsignificant negative coefficient. Finally, as was the case in mathematics in Missouri, we observe no significant effect of the program on students of mentee teachers in reading.

Table 9. Impact of the Leading Educators Fellowship Program on Reading Achievement in Missouri

	Fellows vs. Applicants	Fellow PSM	Fellow Prior Effects PSM	Mentee PSM	Mentee Prior Effects PSM
Fellow	0.092+	0.056+	–0.036		
	(0.048)	(0.029)	(0.072)		
Applicant	–0.025				
	(0.055)				
Mentee teacher				0.023	–0.026
				(0.017)	(0.023)
Application score	–0.004+				
	(0.002)				
Student demographics	Y	Y	Y	Y	Y
School year and grade indicators	Y	Y	Y	Y	Y
Classroom composition	Y	Y	Y	Y	Y
School composition	Y	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y	Y
Observations	210,038	3,734	3,270	26,243	32,176
R-squared	0.66	0.63	0.63	0.63	0.64
N of fellow-years (F)	20	21	15		
N of denied applicant–years (DA)	19				
N of mentee teacher–years (MT)				196	120

NOTE: Standard errors in parentheses, clustered at the teacher-by-year level. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

In Tables 10–13, we present results related to the impact of the program on fellows and mentees in Louisiana, in each of four subjects: mathematics, reading, science, and social studies. Table 10 details results among mathematics teachers in Louisiana, where we observe consistently positive effect estimates across model specifications for both fellows and mentees. While the coefficients varies substantially for fellows, ranging from 0.129 to 0.415 standard deviations across the three model specifications, in each case the estimated program impacts were significant or marginally significant and positive. Among mentees, estimated effects are a marginally significant 0.059 standard deviations in a model that matched mentees on contemporaneous observable characteristics and a nonsignificant 0.070 standard deviations in a model that matched mentees based on their pretreatment effectiveness. It is notable that in Louisiana the sample of mentees that could be matched on pretreatment performance is only a small subset of the total sample of mentees, with 17 teacher-year observations, as compared with 52 teacher-year observations in a contemporaneous PSM model. This pattern of substantially smaller samples for mentees in Louisiana when matching on pretreatment performance is similar in other subject areas, and reflects the fact that a large percentage of mentees in Louisiana were new teachers when they first received mentoring.

Table 10. Impact of the Leading Educators Fellowship Program on Mathematics Achievement in Louisiana

	Fellows vs. Applicants	Fellow PSM	Fellow Prior Effects PSM	Mentee PSM	Mentee Prior Effects PSM
Fellow	0.415*	0.129**	0.212+		
	(0.211)	(0.051)	(0.115)		
Applicant	0.005				
	(0.097)				
Mentee teacher				0.059+	0.070
				(0.034)	(0.057)
Application score	–0.007				
	(0.009)				
Student demographics	Y	Y	Y	Y	Y
School year and grade indicators	Y	Y	Y	Y	Y
Classroom composition	Y	Y	Y	Y	Y
School composition	Y	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y	Y
Observations	199,426	1,994	3,935	17,661	9,018
R-squared	0.60	0.53	0.67	0.51	0.63
N of fellow-years (F)	7	7	5		
N of denied applicant–years (DA)	21				
N of mentee teacher–years (MT)				52	17

NOTE: Standard errors in parentheses, clustered at the teacher-by-year level. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

Corresponding results for reading teachers in Louisiana are shown in Table 11. Here, we observed a significantly negative effect estimate of –0.140 standard deviations among fellows in a specification that matches teachers on pretreatment performance, but no apparent positive or negative effect among fellows in other model specifications. We also observed no significant positive or negative effect for the students of mentee reading teachers in either model specification.

Table 11. Impact of the Leading Educators Fellowship Program on Reading Achievement in Louisiana

	Fellows vs. Applicants	Fellow PSM	Fellow Prior Effects PSM	Mentee PSM	Mentee Prior Effects PSM
Fellow	0.044	0.008	–0.140**		
	(0.075)	(0.028)	(0.050)		
Applicant	0.070				
	(0.051)				
Mentee teacher				0.013	0.046
				(0.027)	(0.056)
Application score	–0.004				
	(0.003)				
Student demographics	Y	Y	Y	Y	Y
School year and grade indicators	Y	Y	Y	Y	Y
Classroom composition	Y	Y	Y	Y	Y
School composition	Y	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y	Y
Observations	269,980	9,007	7,666	16,605	12,757
R-squared	0.60	0.51	0.61	0.53	0.58
N of fellow-years (F)	23	22	18		
N of applicant-years (DA)	34				
N of mentee teacher–years (MT)				72	25

NOTE: Standard errors in parentheses, clustered at the teacher-by-year level. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

Finally, in Tables 12 and 13 we present results for science and social studies teachers in Louisiana, respectively. Among students of fellows in science, we observed a positive and significant program effect in a model specification that matched teachers according to their pretreatment effectiveness. However, the coefficients on our other two model specifications for fellows were smaller and not significant. We observed no significant effects of the program on students of mentee teachers in science in Louisiana. Finally, as shown in Table 13, we observed a significant positive effect of the program on students' social studies achievement in mentees' classrooms in our PSM model specification, with positive but nonsignificant effects observed in a specification that matched mentees according to their pretreatment effectiveness. However, we observe negative effects in social studies for students of fellows, though due to the very small sample of fellows (N=2 teachers), we were unable to effectively match these individuals to comparable peers using the pretreatment effectiveness modeling approach.

Overall, the estimated impact of the program on student achievement appears to have been mixed, with some differences across states. In Louisiana, we observed significant or marginally significant positive program effects on mentees in mathematics and social studies. However, we observed no significant program effects on mentees in reading or science in Louisiana, nor any significant effects on mentees in Missouri. Among fellows, our results vary widely across states, subjects, and model specifications, including both significant negative and positive effect estimates.

Table 12. Impact of the Leading Educators Fellowship Program on Science Achievement in Louisiana

	Fellows vs. Applicants	Fellow PSM	Fellow Prior Effects PSM	Mentee PSM	Mentee Prior Effects PSM
Fellow	0.021	–0.132	0.297+		
	(0.074)	(0.115)	(0.167)		
Applicant	0.050				
	(0.119)				
Mentee teacher				0.064	–0.040
				(0.041)	(0.059)
Application score	0.004				
	(0.006)				
Student demographics	Y	Y	Y	Y	Y
School year and grade indicators	Y	Y	Y	Y	Y
Classroom composition	Y	Y	Y	Y	Y
School composition	Y	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y	Y
Observations	194,112	2,221	5,066	10,475	8,398
R-squared	0.56	0.41	0.59	0.50	0.56
N of fellow-years (F)	9	9			
N of denied applicant–years (DA)	15				
N of mentee teacher–years (MT)			7	37	18

NOTE: Standard errors in parentheses, clustered at the teacher-by-year level. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

Table 13. Impact of the Leading Educators Fellowship Program on Social Studies Achievement in Louisiana

	Fellows vs. Applicants	Fellows PSM	Mentee PSM	Mentee Prior Effects PSM
Fellow	–0.323	–0.284**		
	(0.237)	(0.067)		
Applicant	0.061			
	(0.135)			
Mentee teacher			0.079**	0.048
			(0.040)	(0.101)
Application score	0.004			
	(0.016)			
Student demographics	Y	Y	Y	Y
School year and grade indicators	Y	Y	Y	Y
Classroom composition	Y	Y	Y	Y
School composition	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y
Observations	199,011	1,373	12,362	3,843
R-squared	0.48	0.48	0.40	0.48
N of fellow-years (F)	5	4		
N of applicant-years (A)	16			
N of mentee teacher-years (MT)			44	10

NOTE: Standard errors in parentheses, clustered at the teacher-by-year level. Statistical significance denoted by **p<0.01, *p<0.05, +p<0.10.

Retention

We now turn to the results of our analysis of retention rates among participating program teachers. These include a descriptive analysis of the overall rates at which teachers stayed in the same school, remained in teaching, or remained in any high-poverty school, as well as a regression analysis that compares the likelihood of these retention outcomes for mentees that may have been influenced by the LE Fellowship program, relative to those for an observably similar sample of matched comparison teachers.[28]

In Tables 14 and 15, we provide descriptive information on retention outcomes for teachers in Missouri and Louisiana, respectively. We observed that participating teachers (either fellows or mentees) frequently changed schools, with one-year retention rates between 70 and 86 percent in Missouri and between 50 and 75 percent in Louisiana.[29] In both states, retention rates in the same school and in teaching were somewhat lower than the average for our state-wide sample of teachers. However, program teachers remained in high-poverty schools at rates that were generally comparable to or higher than that of other teachers in our sample. Overall, the differences that we observed between retention rates of participating teachers and other teachers in each state may be related to the types of schools that fellows worked in. We observe that denied applicants—who similarly worked in high-poverty schools—had retention outcomes that were more similar to fellows than to other teachers in the state.

[28] While we were able to obtain data on promotions to principal and central office roles in Missouri, we did not observe any of these promotions among fellows or mentees during the study period.

[29] The Louisiana retention rates do not take into account teachers who are promoted. For example, if a teacher stays in the same school and becomes an assistant principal, this is not captured in the data, and this teacher is considered to have exited from the school. In addition, there were a significant number of school closings and new school openings during the study period.

Table 14. Retention Rates over Time for Teacher Cohorts and Subgroups, Missouri

	Fellows	Denied Applicant	Mentee Teacher	District Teachers
Same school retention				
2011 cohort, 1 year retention	0.69	0.73	0.86	0.83
	26	30	80	67,907
2011 cohort, 2 years retention	0.50	0.47	0.46	0.7
	26	30	80	67,907
2012 cohort, 1 year retention	0.90	0.96	0.77	0.82
	20	30	49	68,005
Teaching retention				
2011 cohort, 1 year retention	0.77	0.90	0.91	0.9
	26	30	77	67,907
2011 cohort, 2 years retention	0.73	0.77	0.86	0.83
	26	30	77	67,907
2012 cohort, 1 year retention	0.95	0.83	0.86	0.9
	20	24	44	68,005
High poverty school retention				
2011 cohort, 1 year retention	0.96	0.97	0.94	0.79
	26	30	80	13,091
2011 cohort, 2 years retention	0.96	0.93	0.9	0.67
	26	30	80	13,091
2012 cohort, 1 year retention	0.95	0.96	0.94	0.76
	20	25	49	13,630

NOTE: Sample of teachers consists of all teachers in districts in Missouri with at least one program applicant in any year. Sample includes all teachers across grades and role designations, including teachers who do not teach students in tested subject areas, and both full- and part-time staff. Subsample cohorts reflect a fixed sample of teachers from the year of their first involvement with the program.

Table 15. Retention Rates and Sample Sizes for Teacher Cohorts and Subgroups, Louisiana

	Fellows	Denied Applicant	Mentee Teacher	District Teachers
Same school retention				
2011 cohort, 1 year retention	0.50	0.51	0.49	0.74
	16	38	43	56,066
2011 cohort, 2 years retention	0.19	0.32	0.28	0.58
	16	38	43	56,066
2012 cohort, 1 year retention	0.58	0.47	0.60	0.72
	19	15	83	54,700
Teaching retention				
2011 cohort, 1 year retention	0.79	0.71	0.80	0.88
	16	38	43	56,066
2011 cohort, 2 years retention	0.63	0.53	0.69	0.77
	16	38	42	55,406
2012 cohort, 1 year retention	0.63	0.87	0.83	0.86
	19	15	83	54,700
High poverty school retention				
2011 cohort, 1 year retention	0.79	0.60	0.65	0.68
	16	35	43	29,765
2011 cohort, 2 years retention	0.63	0.39	0.49	0.53
	16	35	43	29,765
2012 cohort, 1 year retention	0.56	0.40	0.74	0.71
	16	10	70	27,477

NOTE: Sample of teachers consists of all teachers in districts in Missouri with at least one program applicant in any year. Sample includes all teachers across grades and role designations, including teachers who do not teach students in tested subject areas and both full- and part-time staff. Subsample cohorts reflect a fixed sample of teachers from the year of their first involvement with the program.

In Tables 16 and 17, we present results from regression models that predict whether mentees in the LE Fellowship program were more or less likely to remain in the same school, in teaching, or in a high-poverty school.[30] To account for possible differences in teacher context that may have influenced retention rates, in this analysis we compared mentee teachers with other teachers in the state who appeared similar in terms of their own and their students' (and schools') observable characteristics.

[30] Coefficient estimates are odds ratios from logit models. A coefficient less than 1 indicates that mentee teachers are less likely to be retained, whereas a coefficient greater than 1 indicates that mentee teachers are more likely to be retained than the comparison teachers

Table 16. Impact of the Leading Educators Fellowship Program on Teacher Retention, Missouri Mentees

	1-Year School Retention	2-Year School Retention	1-Year High-Poverty Retention	2-Year High-Poverty Retention	1-Year Teaching Retention	2-Year Teaching Retention
Mentee teacher	0.580**	0.357	0.409+	0.211	0.819**	0.772**
	(0.221)	(0.263)	(0.232)	(0.2684)	(0.283)	(0.291)
Student demographics	Y	Y	Y	Y	Y	Y
Time indicators	Y	Y	Y	Y	Y	Y
School composition	Y	Y	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y	Y	Y
Pseudo R-squared	0.11	0.12	0.09	0.09	0.07	0.08
Observations	699	486	643	459	699	454
N of mentee teachers	128	84	122	80	128	83

NOTE: Coefficients are odds ratios from logit models. Standard errors in parentheses, clustered at the teacher level
Statistical significance denoted by **p<0.01, *p<0.05, $^{+}$p<0.10.

Table 17. Impact of the Leading Educators Fellowship Program on Teacher Retention, Louisiana Mentees

	1-Year School Retention	2-Year School Retention	1-Year High-Poverty Retention	2-Year High-Poverty Retention	1-Year Teaching Retention	2-Year Teaching Retention
Mentee teacher	0.834	0.493+	1.164	0.675	1.600+	0.840
	(0.192)	(0.196)	(0.305)	(0.265)	(0.425)	(0.345)
Student demographics	Y	Y	Y	Y	Y	Y
Time indicators	Y	Y	Y	Y	Y	Y
School composition	Y	Y	Y	Y	Y	Y
Teacher characteristics	Y	Y	Y	Y	Y	Y
Pseudo R-squared	0.16	0.25	0.15	0.17	0.11	0.17
Observations	857	363	752	327	892	358
N of mentee teachers	107	42	110	42	110	42

NOTE: Coefficients are odds ratios from logit model. Standard errors in parentheses, clustered at the teacher level
Statistical significance denoted by **p<0.01, *p<0.05, $^{+}$p<0.10

The results of these analyses were inconsistent across states and cohorts of mentees. For instance, in Missouri, we observed that mentees were significantly less likely to stay in the same school, in teaching, and in high-poverty schools compared with similar teachers. However, in Louisiana we observed that mentees were more likely to stay in teaching after one year, though these effects were only marginally significant. These varied results were inconsistent with either a uniformly positive or negative effect of the LE Fellowship program on mentees' retention rates.

Overall, our results offer evidence of several instance of large effect sizes associated with teachers' exposure to the LE Fellowship program, either as fellows or as mentee teachers. At the same time, however, we did not observe impacts on student achievement across all state-subject area combinations, and the results of our retention analyses are mixed. In the following section, we provide concluding thoughts regarding both the potential implications and limitations of these results.

7. Conclusions and Discussion

Little research to date has examined the impacts of developing mid-career teachers as teacher leaders and mentors in schools. While we were unable to conduct an ideal experiment, in this study we used rich administrative data on students, teachers, and schools and employed a range of rigorous analytical techniques, including quasi-experimental methods endorsed by the federal What Works Clearinghouse, to provide a rigorous evaluation of possible effects of the LE Fellowship program. Although our sample size and resulting power for detecting program effects were limited, this report nevertheless constitutes valuable new evidence regarding the potential impacts of a growing trend in teacher development efforts.

Our preliminary evaluation of program effects on the student achievement gains of participating fellows revealed mixed and inconsistent findings. The effects that we detect range in size and include both positive and negative effect sizes corresponding to changes of as much as 10 percentage points in student achievement rankings. In Missouri, we observed marginally significant and positive effects among fellows who teach reading in some modeling specifications, but a large and significant negative effect in one model specification among fellows who teach math. In Louisiana, we observe some significant positive effects among fellows who teach math and science, but a significant negative effect among fellows who teach reading and social studies. With the exception of the apparent positive impacts among fellows who teach math in Louisiana, these findings are generally inconsistent across different modeling specifications. Moreover, the available sample size of fellows is often quite small, which means that we may lack statistical power to detect true program effects in some cases.

The statistically significant effect sizes that we did observe for mathematics achievement in Louisiana were substantial when compared with other prominent benchmarks in the field of education policy. For instance, considering the most conservative estimate of 0.059 standard deviations for mentees, the differential positive effect of these teachers is equivalent to roughly a one-half standard deviation difference in teacher "value added" effectiveness (Rockoff, 2004). The estimated differential effect of program fellows is even larger, with the conservative estimate of 0.129 standard deviations equal to approximately 40 percent of the benefit that students experience from attending a highly effective urban charter school (Betts and Tang, 2014; Angrist, Pathak, and Walters, 2011).

Overall, we recommend caution in interpreting the findings with respect to program impacts on fellows, and we note that the results could be explained by a variety of mechanisms. First, the sample of fellows analyzed in these models is restricted to the small subset of fellows who are teachers of record and can be directly linked to students. Some of these fellows may benefit from the program training and from increased interaction and knowledge-sharing with mentee teachers, and this improves their instruction. Alternatively, some of these fellows may have

reduced instructional effectiveness, potentially because they did not receive enough release time for their expanded responsibilities as instructional leaders. However, given the inconsistency of results across model specifications and the relatively large size of some estimates, it is also possible that our methods reflect a biased comparison between treated and comparison teachers as a function of unobserved, preexisting differences between fellows and other teachers. Additional evidence from future reports can help to better flesh out these competing hypotheses, and in future reports we may have sufficient sample sizes to separate out program effects during and subsequent to the LE training.

Among mentee teachers, where we have larger sample sizes, our analyses constitute a more appropriately powered evaluation of possible program impacts. Here, we observe no significant positive or negative effects of the program among mentees in Missouri. However, we do see some suggestive evidence of beneficial program impacts among mentees in Louisiana, in two of four subject areas. Our estimated effect sizes in these subjects correspond to roughly a 2–3 percentage point increase in students' achievement rankings from one year to the next. However, it is important to note that we are investigating multiple subject areas and that the statistical significance of these estimates are not robust to an adjustment for multiple hypotheses. Future analyses that include a larger sample of mentees in Louisiana will be better powered to verify the possible beneficial impacts of the program in that state. In addition, we may also have sufficient sample in both states in the future to better differentiate between any immediate, versus longer-term, effects of being mentored by fellows.

In addition to effects on student achievement gains, we evaluated whether the LE Fellowship program may have impacted teacher retention. In this area, however, we failed to identify any consistent positive or negative effect of the program on participating mentee teachers. One interpretation of the sporadic and in some cases divergent impacts that we observed could be that the program varied in its impact on teacher retention across cohorts, years, and state contexts. Alternatively, and perhaps more likely, we may have been unable to fully account for all of the key drivers of teacher retention in our analyses, in spite of employing a range observable student, teacher, and school characteristics. LE Fellowship program participants may have differed from their peers in important unobserved ways with respect to their career trajectories and ambitions, either overall or dynamically across different cohorts and states.

Overall, this study provides new insights into the potential benefits of developing mid-career teacher leaders. Future analyses with larger samples of both fellows and mentees will provide greater statistical power for more reliably detecting any true program effects on student achievement and teacher retention, and may also allow for a more nuanced investigation of program effects both during and subsequent to participation in the program.

Technical Appendix

Comparison of Fellows and Denied Applicants

Our first specification for evaluating the differential performance of LE fellows relative to denied applicants was as follows:

$$A_{isgjt} = A_{isgj(t-1)}\beta_1 + X_{it}\beta_2 + C_{ijt}\beta_3 + S_{ist}\beta_4 + F_{ijt}\beta_5 + FA_{ijt}\beta_6 + App_{ij}\beta_7 + Int_{ij}\beta_8 + T_{ijt}\beta_9 + \theta_t + \mu_g + \epsilon_{isgjt} \quad (1)$$

For each subject area of interest (mathematics and communications), we estimated achievement A_{isgjt} for student *i* at school *s* in grade *g* with teacher *j* during year *t* as a function of prior-year achievement in the subject ($A_{isgj(t-1)}$); a vector X of observable student characteristics including race, gender, free-lunch status, LEP status, special education status, and an indicator of time spent in the school building; a vector C of classroom-level characteristics; and a vector S of school-level characteristics, analogous to the student-level information aggregated to the classroom or school attended by student *i* in year *t*. School characteristics also include indicators for whether the school is a charter school or a Title 1 school. Finally, we also include a vector T of time-varying and time-invariant teacher characteristics including gender, race, salary, degree earned, and indicators for years of teaching experience and a series of fixed effects for school years θ_t and grade levels μ_g.

The key coefficient of interest here is β_5, the coefficient on the indicator variable F_{ijt}, which is equal to 1 in all years after a teacher *j* of student *i* is admitted into the LE Fellowship program, and 0 otherwise. To facilitate a comparison between students of program applicants who were accepted and those who were denied acceptance, we also include an indicator FA_{ijt} that is equal to 1 in all periods after a teacher applies to the program (regardless of whether or not they are accepted). Thus, β_4 represents the differential achievement gains of students of program participants, over and above that of students of applicants. Note that our sample also includes teachers who are neither denied applicants or fellows, as this allows us to control for the full range of observable student, classroom, and school characteristics that are included in all of our models.

To control for possible differences in the performance of students of accepted and rejected applicants who are driven by the applicant screening process, we include a continuous, time-invariant variable App_{ij}, which is the overall application score used to screen and determine admission into the LE Fellowship program. For the same reason, we also include an indicator, Int_{ij}, that is set to 1 for those applicants who received a failing score on any individual screening interview, as this was sometimes an additional criteria used to reject applicants. Both the

application score and low-interview indicator controls are set to zero for teachers who did not apply to the program. Standard errors are clustered at the teacher-by-year level to account for the fact that errors may be correlated for a given class and subject.

Propensity Score Matching and Regression

In our PSM approach, we did not utilize data on applicants to the program to compare the impacts of successful applicants relative to denied applicants. Instead, we used a probit model to predict the likelihood of each treatment (being a fellow or a mentee) in order to identify a matched comparison group, and then specified a linear regression model to estimate differences in adjusted outcomes between the treated and comparison groups.

In particular, we estimated teachers' likelihood of being treated or not (in a given year) as a function of a vector of teacher-level characteristics *T* for teacher *j* at time *t*; a vector of their classroom characteristics C; a vector of school characteristics S; and year fixed effects θ_t. Each of the vectors *T*, *C*, and *S* included the same variables as the corresponding vectors specified in the description of the fellow and applicant comparison model above:

$$L^* = T_{jt}\beta_1 + C_{jt}\beta_2 + S_{jt}\beta_3 + \theta_t + \epsilon_{jt} \quad (2)$$

We estimated separate interest propensity scores for fellows and mentees. However, due to our limited sample of fellows, matching models for fellows include a single indicator for teachers' race (i.e., white/nonwhite), and a single indicator for teachers' degree level (master's or higher versus not).

As described in Section 5 of this report, we utilized many-to-one nearest-neighbor matching with replacement to identify up to ten matches for treated observations at the teacher-year level, as appropriate to our research questions. We restricted all matches to be within a propensity score range of 0.01 or less to the treated observation in question, which corresponds to, at most, a 0.5 standard deviation difference in observed propensity scores across our teacher and school samples. Across all subject areas and across analyses of participating fellows or mentees, the region of common support included both treated observations and at least some potential matches with similar likelihood of treatment. Our matching models varied in their explanatory power, with adjusted R-squared ranging from around 0.15 to 0.40.

Subsequent to matching, we pooled our treated and comparison samples and estimated the effects of treatment status on student achievement outcomes in each year:

$$A_{isgjt} = A_{isgj(t-1)}\beta_1 + X_{it}\beta_2 + C_{ijt}\beta_3 + S_{ist}\beta_4 + T_{ijt}\beta_5 + M_{ijt}\beta_6 + \theta_t + \mu_g + \epsilon_{isgjt} \quad (3)$$

Here, our notation and specification is very similar to that shown for Equation 1 above. However, this model does not include controls for denied applicants or for applicant screening

rating scores. Instead, we relied on the propensity matching to facilitate a direct comparison between comparable treated and control teachers who appeared equally likely to have participated in the fellowship program. Here, β_6 represents the difference in effectiveness of teachers who participated in the program (either fellows or mentees, respectively) relative to the matched control group. We continued to include controls for observable student (*X*), classroom (*C*), teacher (*T*), and school-level (*S*) characteristics in order to adjust for any lingering differences between treatment and matched comparison schools, while also improving the precision of our estimates.

Matching on Prior-Treatment Instructional Effectiveness

As an alternative to the previous modeling approaches, we instead, for a subsample of teachers who we could observe in the year prior to receiving treatment, matched treated fellows or mentees to peer teachers in order to ensure that they were equivalent in a baseline (i.e., prior) school year in terms of their measured effects on students' achievement gains. To facilitate this matching, we estimated teacher-by-year effects across our sample via a two-step approach. First, we estimated individual student achievement gains, accounting for a range of observable characteristics of each student and of their class and school peers that may have influenced contemporaneous achievement, as detailed in Equation 3:

$$A_{isgjt} = A_{isgj(t-1)}\beta_1 + X_{it}\beta_2 + S_{ist}\beta_3 + C_{ijt}\beta_4 + \theta_t + \mu_g + \epsilon_{isgjt} \quad (4)$$

In this first step, we estimated achievement *A* for student *i* at school *s* in grade *g* with teacher *j* in year *t* as a function of prior-year achievement in the subject in the subject ($A_{(t-1)}$), a vector *X* of observable time-varying student characteristics including race, gender, free-lunch status, LEP status, special education status, and an indicator of time spent in the school building, and vectors C and S of class-level and school-level characteristics, analogous to the student-level information aggregated to the class or school attended by student *i* in year *t*, and a series of fixed effects for school years θ_t and grade levels μ_g.

In the second step, we aggregated the residual variation in student achievement in time *t* that was not explained by these student, class, or school factors, and computed the average of the student-level residuals for each teacher-year observation linked to them in time *t*. We assumed that this unexplained variance corresponds to teacher-by-year contributions to students' measured achievement. We standardized the resulting teacher-by-year measures to have a mean of 0 and a standard deviation of 1.

Next, we matched treated fellow or mentee teachers (separately), by cohort, with potential control teachers. For each treated teacher, we identified control teacher(s) from the year prior to their first year in the program by matching them to teachers who teach students in the same grade level and whose prior-year instructional impacts predicted a similar likelihood of receiving

treatment. We matched each teacher with up to ten "nearest neighbor" matches who are within a caliper of 0.01 in terms of their measured propensity to receive treatment. Our sample of potential control teachers included only teachers who taught at least five students and were present in our sample in both the pretreatment year and in the first year in the treated teacher is treated. Prior effectiveness was not necessarily a strong predictor of treatment (though in most cases it was a significant predictor), but as shown in Appendix A Table A5, this matching successfully identified teachers who were very similar in terms of pretreatment instructional effectiveness.

Finally, we compared the effects of identified treated and control teachers using student-level models (with errors clustered at the teacher-year level) in the same way as for the PSM modeling approach above. In other words, in addition to matching teachers according to their measured effectiveness in the baseline year, we estimated student achievement differences between treated and control teachers, while controlling for the same student, teacher, and school-level controls as in all of our models.

Unlike in our PSM approach where we matched at the level of teacher-year observations, in this modeling approach we identified, for each cohort of treated fellows or mentees, a fixed sample of control teachers with comparable baseline impacts. We then included observations of both treated and control teachers in all years subsequent to their entering the LE Fellowship program. Among both treated and control teachers, we observed some attrition from the sample of teachers for whom we had data after the first year of treatment. Because attrition could be the result of the LE Fellowship program and this might color our results, we also considered the results from a specification check that considered only data from treated and control teachers in the year in which they were first treated by the LE Fellowship program. While this specification likely understated any cumulative impact of participation in the program, we observed that estimated program effects were similar to those presented in our main results section. These results are available upon request.

Alternative Specification Checks

In preliminary analyses, we conducted a number of specification checks. For example, we explored including controls for both mathematics and reading prior-year test scores in the mathematics and reading regressions, but this did not substantially alter our results. Also, we considered specifications that examine treatment effects over time in each year following treatment, but the by-year samples of treated teachers were too small and the results were not significantly different from those presented here. Similarly, our results are consistent when fellow and mentee teacher indicators are included simultaneously in the same regression, versus in separate regressions.

We also explored a modification to Equation 1 above in which we considered the relative within-school performance of treated versus control teachers. We did so by replacing time-

invariant school-level characteristics S with a vector of school fixed effects. If our controls for school characteristics do not fully control for the quality of the school where the teacher works, then the effect of the program may be biased. However, under the alternative specification that uses school fixed effects, our results could be biased if there are positive spillover effects of the LE Fellowship program. In any case, we found that results using a school fixed effects model specification were consistently very similar to those that used school-characteristics controls.

Finally, we examined specifications whether there were any school spillover effects by estimating the impact of the school having a fellow on student achievement gains. There were no statistically significant results to indicate that there were school spillover effects.

Teacher Retention

To assess whether participation in the LE Fellowship program spurs changes in teacher retention rates, we first matched each participating LE teacher (either fellow or mentee) to a "control" teacher, in the year in which they first enter the program. As in the PSM analyses for student achievement, we matched teachers according to their observed probability of receiving treatment in that year. As before, we matched according to our full range of available student, teacher, and school-wide characteristics that may predict entry into the program. We matched each treated teacher with up to ten "nearest neighbor" control teachers, and use a caliper restriction of 0.02 for their observed propensity to be fellows or mentees, respectively.

Next we used logistic regressions to estimate whether participation in the program predicted differences in teacher retention, when controlling for other observables that might also influence retention:

$$P(R_{jst}) = \beta_0 + T_{jt}\beta_1 + S_{jst}\beta_2 + Ment_j\beta_3 + \theta_t + \upsilon_{sgjt} \qquad (5)$$

For each retention outcome (staying in the same school, a high-poverty school, or teaching in the state), we estimated the probability of retention for teacher *j* at school *s* during year *t* as a function of a vector T of time-varying and time-invariant teacher characteristics including gender, race, salary, degree earned, and indicators for years of teaching experience, a vector S of school-level characteristics with student-level information aggregated to the school including race, gender, free-lunch status, LEP status, special education status, and an indicator for charter schools, and a series of fixed effects for school years θ_t. We assumed that the error term is normally distributed.

The key coefficient of interest is β_3, the coefficient on the indicator variable for mentee teacher, which is equal to 1 if teacher *j* was a mentee teacher in that year, and 0 otherwise. Standard errors were clustered at the teacher-by-year level.

Appendix A. Propensity Score Matching and Power Analyses

Table A1. Standard Deviation Differences on Observable Characteristics Between Treated Mentees in Missouri, PSM-Matched Comparison Teachers, and the Initial Total Sample of Teachers

	Mathematics		Reading	
	Diff. from Matches (T – M)	***Diff. from Sample (T – S)***	***Diff. from Matches (T – M)***	***Diff. from Sample (T – S)***
Teacher characteristics				
Salary	0.075	–0.336	0.029	–0.304
Female	–0.049	–0.082	0.064	–0.112
White	–0.033	–0.499	–0.040	–0.639
1st-year teacher	–0.023	0.135	–0.029	0.108
2nd-year teacher	–0.133	0.212	–0.013	0.150
3rd-year teacher	0.116	0.336	0.052	0.221
4th-year teacher	–0.029	0.114	–0.018	0.145
5th-year teacher	–0.004	–0.035	–0.011	0.037
6th-year teacher	0.019	–0.128	0.045	–0.005
7th-year teacher	0.037	–0.101	0.003	–0.118
8th-year teacher	0.009	–0.066	–0.019	–0.034
9th-year teacher	–0.014	–0.053	–0.026	–0.077
10th-year teacher or higher	0.031	0.023	–0.026	–0.021
Master's or doctorate degree	0.003	–0.657	0.002	–0.659
School characteristics				
Charter school	–0.061	0.281	–0.044	0.234
School % free/reduced lunch	0.046	0.943	0.003	0.940
School % minority	0.040	1.162	–0.004	1.209
School % female	0.044	–0.100	–0.012	–0.082
School % ESL	0.140	0.709	0.159	0.606
School % special education	–0.203	0.435	–0.048	0.365
School average pretreatment test score	–0.001	–0.764	–0.012	–0.870
Teachers' classroom characteristics				
Classroom % free/reduced lunch	0.006	0.899	–0.007	0.900
Classroom % minority	0.020	0.912	–0.002	0.928
Classroom % female	–0.022	0.010	–0.047	0.026
Classroom % ESL	0.140	0.685	0.052	0.529
Classroom % special education	0.008	0.054	0.004	–0.018
Grade levels taught				
Grade 4 teacher	0.091	0.187	–0.007	0.044
Grade 5 teacher	–0.038	–0.106	–0.009	–0.170
Grade 6 teacher	–0.033	–0.046	0.084	0.090
Grade 7 teacher	0.040	–0.207	0.017	0.106
N of treated teacher-years	159		196	
N of matched teacher-years	689		708	

Table A2. Standard Deviation Differences on Observable Characteristics Between Treated Fellows in Missouri, PSM-Matched Comparison Teachers, and the Initial Total Sample of Teachers

	Mathematics		Reading	
	Diff. from Matches (T – M)	***Diff. from Sample (T – S)***	***Diff. from Matches (T – M)***	***Diff. from Sample (T – S)***
Teacher characteristics				
Salary	–0.135	0.212	0.015	0.077
Female	–0.060	–0.101	0.007	0.266
White	–0.007	–1.192	0.152	–1.407
1st-year teacher	0.000	–0.310	0.000	–0.308
2nd-year teacher	0.013	–0.143	–0.032	–0.111
3rd-year teacher	–0.035	0.280	0.203	0.582
4th-year teacher	0.157	0.218	0.000	–0.210
5th-year teacher	0.032	0.142	0.017	–0.024
6th-year teacher	–0.011	0.182	–0.159	0.003
7th-year teacher	–0.102	0.175	–0.193	0.013
8th-year teacher	0.040	–0.017	0.131	0.261
9th-year teacher	0.000	–0.209	0.000	–0.206
10th-year teacher or higher	0.000	–0.203	0.000	–0.204
Master's or doctorate degree	–0.060	–0.420	0.069	–0.414
School characteristics				
Charter school	0.161	0.458	–0.003	0.118
School % free/reduced lunch	0.042	1.092	–0.008	1.026
School % minority	0.017	1.143	–0.007	1.153
School % female	–0.043	0.152	0.053	–0.134
School % ESL	0.388	1.911	0.672	1.910
School % special education	–0.040	–0.473	–0.206	–0.457
School average pretreatment test score	–0.052	–0.444	0.146	–0.693
Teachers' classroom characteristics				
Classroom % free/reduced lunch	0.044	1.031	0.012	0.958
Classroom % minority	0.006	0.943	–0.031	0.927
Classroom % female	–0.115	0.380	–0.013	0.228
Classroom % ESL	0.368	1.993	0.574	1.620
Classroom % special education	–0.073	–0.329	0.079	–0.238
Grade levels taught				
Grade 4 teacher	0.106	0.164	0.270	0.332
Grade 5 teacher	–0.168	–0.150	–0.060	–0.141
Grade 6 teacher	–0.061	–0.100	–0.021	–0.210
Grade 7 teacher	0.170	0.119	–0.227	0.123
N of treated teacher-years	24		21	
N of matched teacher-years	131		118	

Table A3. Standard Deviation Differences on Observable Characteristics Between Treated Mentees in Louisiana, PSM-Matched Comparison Teachers, and the Initial Total Sample of Teachers

	Mathematics		Reading		Science		Social Studies	
	Diff. from Matches (T – M)	*Diff. from Sample (T – S)*	*Diff. from Matches (T – M)*	*Diff. from Sample (T – S)*	*Diff. from Matches (T – M)*	*Diff. from Sample (T – S)*	*Diff. from Matches (T – M)*	*Diff. from Sample (T – S)*
Teacher characteristics								
Female	–0.041	–0.338	0.064	–0.155	–0.088	–0.286	0.163	–0.122
Asian	0.000	–0.143	–0.005	–0.017	0.000	–0.138	0.000	–0.102
Black	–0.052	0.323	0.120	–0.070	0.008	0.067	0.017	0.088
American Indian	0.000	0.000	0.000	–0.045	0.000	–0.038	0.000	0.000
Hispanic	0.000	–0.119	–0.050	–0.004	0.000	–0.118	0.092	0.067
Pacific Islander	0.000	–0.024	0.000	–0.014	0.000	–0.019	0.000	–0.018
Multi-racial	0.000	–0.038	0.000	–0.043	0.000	–0.053	0.000	–0.040
Master's degree	0.033	–0.184	–0.042	–0.075	0.032	–0.266	0.051	–0.146
Master's plus 30 credits	–0.053	–0.118	0.000	–0.234	0.000	–0.216	0.000	–0.209
Doctorate degree	0.122	0.266	0.021	0.155	0.492	0.669	0.241	0.551
Specialist degree	0.000	–0.077	0.000	–0.083	0.000	–0.071	0.000	–0.079
1st-year teacher	0.038	0.366	0.001	0.382	0.077	0.350	0.068	0.025
Estimated years of experience >1 (continuous)	–0.055	–0.245	0.100	0.051	0.044	–0.105	0.068	0.025
School characteristics								
Charter school	0.009	0.882	0.146	1.521	0.141	1.360	0.069	1.428
Title 1 schools	–0.168	–0.128	–0.302	0.594	0.070	–0.115	–0.115	0.064
School % free/reduced lunch	–0.024	0.673	0.054	0.722	0.041	0.680	0.010	0.766
School % minority	–0.009	0.670	0.079	0.540	0.010	0.765	0.021	0.640
School % female	0.129	0.025	0.123	–0.306	0.080	0.017	0.188	–0.175
School % ESL	0.020	–0.156	0.004	–0.420	–0.087	–0.197	–0.096	–0.306
School % special education	–0.214	0.676	–0.138	0.840	–0.047	0.791	–0.178	0.878
School % highly mobile	0.647	1.144	0.017	1.125	1.326	1.637	0.503	1.261
School average pretreatment test score	0.041	–0.516	0.030	–0.613	–0.025	–0.529	0.051	–0.753
Teachers' classroom characteristics								
Classroom % free/reduced lunch	–0.044	0.622	0.035	0.674	–0.014	0.614	0.004	0.737
Classroom % minority	–0.016	0.664	0.085	0.531	–0.001	0.787	0.009	0.653
Classroom % female	0.028	0.064	0.071	–0.185	0.047	–0.144	0.087	–0.238
Classroom % ESL	–0.007	–0.159	0.008	–0.266	–0.040	–0.225	–0.040	–0.287
Classroom % special education	–0.074	0.036	–0.047	0.269	0.122	0.723	–0.099	0.641
Classroom % highly mobile	0.516	0.920	–0.066	0.926	1.123	1.620	0.435	1.055
Grade levels taught								
Grade 4 teacher	0.162	–0.056	0.062	0.041	0.047	0.001	0.119	–0.106
Grade 5 teacher	0.074	–0.082	–0.055	–0.185	–0.010	–0.329	0.007	–0.059
Grade 6 teacher	–0.101	0.069	0.080	0.154	–0.020	0.479	–0.034	0.179
Grade 7 teacher	–0.137	0.088	–0.058	0.060	–0.098	0.000	–0.075	–0.016
Grade 8 teacher	–0.069	0.019	–0.044	–0.052	0.069	–0.067	–0.077	0.064
N of treated teacher-years	52		72		37		44	
N of matched teacher-years	338		387		186		232	

Table A4. Standard Deviation Differences on Observable Characteristics Between Treated Fellows in Louisiana, PSM-Matched Comparison Teachers, and the Initial Total Sample of Teachers

	Mathematics		Reading		Science		Social Studies	
	Diff. from Matches (T – M)	***Diff. from Sample (T – S)***	***Diff. from Matches (T – M)***	***Diff. from Sample (T – S)***	***Diff. from Matches (T – M)***	***Diff. from Sample (T – S)***	***Diff. from Matches (T – M)***	***Diff. from Sample (T – S)***
Female	0.070	–0.024	–0.234	–0.017	0.000	–1.496	–0.618	–0.571
White	0.000	0.759	0.150	0.295	0.256	–0.413	0.000	0.698
Master's or doctorate degree	0.005	0.022	–0.013	–0.103	0.246	–0.166	–0.331	–0.198
1st-year teacher	0.000	–0.562	0.066	–0.349	0.000	–0.637	0.000	–0.195
Estimated years of experience >1 (continuous)	–0.230	–0.356	0.083	–0.071	–0.114	–0.061	–0.133	–0.371
School % free/reduced lunch	0.044	0.849	0.008	0.735	–0.034	0.865	0.026	0.834
Charter school	–0.071	1.872	–0.030	1.594	0.000	2.198	0.208	1.798
Charter school x % free/reduced lunch	–0.044	1.925	–0.025	1.645	–0.019	2.258	0.227	1.859
Teacher % free/reduced lunch	0.022	0.775	–0.011	0.668	–0.032	0.824	0.130	0.675
School average pretreatment test score	–0.189	–0.699	–0.042	–0.371	0.022	–0.754	–0.101	–0.218
Grade 4 teacher	0.132	0.213	–0.032	–0.134	0.133	–0.200	–0.364	0.079
Grade 5 teacher	–0.017	–0.233	0.233	0.039	–0.298	–0.127	0.169	–0.159
Grade 6 teacher	0.290	0.397	–0.060	–0.204	–0.322	0.214	0.000	–0.400
Grade 7 teacher	0.000	–0.402	–0.082	–0.037	0.000	–0.389	0.624	0.262
Grade 8 teacher	–0.447	–0.004	–0.104	0.402	0.557	0.616	–0.316	0.275
N of treated teacher-years	7		23		9		5	
N of matched teacher-years	43		171		36		34	

Table A5. Standardized Differences in Instructional Efficacy Prior to Treatment Between Treated Teachers, Matched Comparison Teachers, and the Initial Total Sample

	Diff. from Matches (T – M)	Diff. from Sample (T – S)	N of Teachers Treated	N of Teachers Matched
Mathematics				
Missouri mentees	–0.003	0.173	52	420
Missouri fellows	–0.002	0.177	10	100
Louisiana mentees	–0.004	0.115	13	130
Louisiana fellows	–0.012	–0.516	4	40
Reading				
Missouri mentees	0.000	0.081	62	526
Missouri fellows	0.002	0.578	9	90
Louisiana mentees	0.005	0.038	19	181
Louisiana fellows	0.005	0.497	11	100
Science				
Louisiana mentees	–0.017	–0.136	10	85
Louisiana fellows	0.031	0.411	5	41
Social studies				
Louisiana mentees	0.081	0.655	8	70
Louisiana fellows	n/a	n/a	2	n/a

Table A6. Power Analysis for Hypothetical Analyses Describing Minimum Detectable Effect Sizes for Various Teacher-Year Sample Sizes

Samples		Assumptions				
N of Treated Observations	N of Control Observations	Alpha	Power	ICC	Level-2 R^2	MDE
5	50	0.05	0.8	0.15	0.5	0.375
10	100	0.05	0.8	0.15	0.5	0.261
20	200	0.05	0.8	0.15	0.5	0.183
40	400	0.05	0.8	0.15	0.5	0.129
80	800	0.05	0.8	0.15	0.5	0.091
160	1,600	0.05	0.8	0.15	0.5	0.065
200	2,000	0.05	0.8	0.15	0.5	0.058
5	50	0.05	0.5	0.15	0.5	0.236
10	100	0.05	0.5	0.15	0.5	0.164
20	200	0.05	0.5	0.15	0.5	0.115
40	400	0.05	0.5	0.15	0.5	0.081
80	800	0.05	0.5	0.15	0.5	0.057
160	1,600	0.05	0.5	0.15	0.5	0.041
200	2,000	0.05	0.5	0.15	0.5	0.036

Abbreviations

LCR	Leadership Competency Rubric
LE	Leading Educators
LEP	limited English proficiency/limited English proficient
MDE	minimum detectable effect size
PSM	propensity score matching

References

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters, "Explaining Charter School Effectiveness," Cambridge, Mass.: National Bureau of Economic Research, NBER Working Paper No. 17332, 2011.

Betts, Julian R., and Y. Emily Tang, *A Meta-Analysis of the Literature on the Effect of Charter Schools on Student Achievement*, Seattle, Wash.: Center on Reinventing Public Education University of Washington, 2014.

Caliendo, Marco, and Sabine Kopeinig, "Some Practical Guidance for the Implementation of Propensity Score Matching," *Journal of Economic Surveys*, Vol. 22, No. 1, 2008, pp. 31–72.

Campbell, Patricia F., and Nathanial N. Malkus, "The Impact of Elementary Mathematics Coaches on Student Achievement," *Elementary School Journal*, Vol. 111, No. 3, 2011, pp. 430–454.

Cohen, B., and E. Fuller, "Effects of Mentoring and Induction on Beginning Teacher Retention," paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., 2006.

Dehejia, Rajeev H., and Sadek Wahba, "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *Review of Economics and Statistics*, Vol. 84, No. 1, 2002, pp. 151–161.

Ferguson, Barbara, "Closing Schools, Opening Schools and Changing School Codes: Instability in the New Orleans Recovery School District," Research on Forums, Inc., June 2014. As of July 27, 2015: www.researchonreforms.org/html/documents/RSDClosingOpeningChangingCodes.pdf

Fuller, E., "Beginning Teacher Retention Rates for TxBESS and Non-TxBESS Teachers," unpublished paper, State Board for Educator Certification, Texas, 2003.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Kazuaki Jones, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, Laura Szteejnberg, 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, NCEE 2008-4030, 2008.

Garet, Michael S., Andrew J. Wayne, Fran Stancavage, James Taylor, Marian Eaton, Kirk Walters, Mengli Song, Seth Brown, Steven Hurlburt, Pei Zhu, Susan Sepanik, Fred Doolittle, and Elizabeth Warner, *Middle School Mathematics Professional Development Impact Study:*

Findings After the Second Year of Implementation, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, 2011.

Glazerman, Steven, Eric Isenberg, Sarah Dolfin, Martha Bleeker, Amy Johnson, Mary Grider, and Matthew Jacobus, *Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study*, Washington, D.C.: U.S. Department of Education, NCEE 2010-4027, 2010.

Ingersoll, Richard, and Thomas M. Smith, "Do Teacher Induction and Mentoring Matter?" *NASSP Bulletin*, Vol. 88, No. 638, 2004a, pp. 28–40.

———, "The Impact of Induction and Mentoring on Beginning Teacher Turnover in High and Low Poverty Schools," paper presented at the annual meeting of the American Educational Research Association, San Diego, Calif., 2004b.

Ingersoll, Richard, and Michael Strong, "The Impact of Induction and Mentoring Programs for Beginning Teachers: A Critical Review of the Research," *Review of Educational Research*, Vol. 81, No. 2, 2011, pp. 201–233.

Jackson, C. Kirabo, and Elias Bruegmann, *Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 15202, 2009.

Koedel, Cory, "An Empirical Analysis of Teacher Spillover Effects in Secondary School," *Economics of Education Review*, Vol. 28*,* No. 6, 2009, pp. 682–692.

McCaffrey, Daniel, F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly, "The Intertemporal Variability of Teacher Effect Estimates," *Education*, Vol. 4, No. 4, 2009, pp. 572–606.

Rockoff, Jonah E., "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, Vol. 94, No, 2, 2004, pp. 247–252.

———, *Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 13868, 2008.

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.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest, REL 2007-No. 033, October 2007.

Yuan, Kun, "A Value-Added Study of Teacher Spillover Effects Across Four Core Subjects in Middle Schools," *Education Policy Analysis Archives*, Vol. 23, No. 38, 2015.