IES Institute of Education Sciences

APPENDIXES
December 2022

Regional Educational Laboratory Southwest
At American Institutes for Research

The Louisiana Believe and Prepare Educator Preparation Reform: Findings from the Pilot and Early Implementation Years

See https://ies.ed.gov/ncee/rel/Products/Publication/104928 for the full report.

Appendix A. Data and methods

This appendix provides additional details about data sources, data preparation, samples, data analysis, sensitivity and robustness checks, and limitations.

Data sources

The study used extant longitudinal administrative data at the teacher and student levels provided by the Louisiana Department of Education (table A1). The data came from seven linkable data collections to allow tracking of teachers and their program completion (including whether the program had implemented Believe and Prepare), certification, and employment outcomes, as well as retention and performance among Believe and Prepare teachers and other early career teachers. The study team used the following two data sources to answer all research questions:

- *The Mentor and Resident Data (MRD) collection,* which includes data on all preservice participants in the full-year residency component of Believe and Prepare, including participants' teacher preparation program, school of residency, residency start and end dates, and certification areas.
- *The Profile of Educational Personnel (PEP) collection,* which includes annual, anonymized records for all teachers employed in Louisiana public schools, including all traditional public schools and charter schools. These records contain information on teacher demographic characteristics, teaching certificates, college degrees, teaching experience, school assignments, district hire dates, teacher mobility across public schools within the state, and teacher exits.

The study team used several additional data sources to answer research questions 1, 3, 4a, and 4c:

- *The Compass Information System (CIS),* which includes teacher performance ratings, including value-added model ratings, student growth ratings, and professional practice ratings from Louisiana's Compass teacher evaluation system for all teachers employed in Louisiana public schools.
- *The Student Information System (SIS),* which includes information on student enrollment, achievement, demographic characteristics, and education needs. Achievement tests include English language arts, math,

social studies, and science for grades 3–8 and end-of-course tests in English I, II, and III; algebra I and geometry; U.S. history; and biology. The analysis was restricted to math and English language arts, subjects for which repeated measures are available as controls for a student's prior achievement.

- *The Curriculum System (CUR),* which collects information on courses and student class schedules, which allows students and teachers to be linked.
- *The Teacher Certification Management System (TCMS),* which collects information on Praxis licensure test scores for both teacher candidates and teachers serving as mentors, certificate type issued (endorsement area and pathway to certification), certificate issue date, start and end dates of certificate validity, and institutions or programs recommending certification.
- *The Teacher Shortage Areas (TSA) website,* which contains information on teacher shortage areas. States report to the U.S. Department of Education the academic disciplines and geographic areas where they are experiencing teacher shortages. These data are publicly available at https://tsa.ed.gov/#/home/ for all school years in the study.

Specifically, teacher performance ratings from CIS were used to address research question 1 (in-service performance ratings), student- and course-level data (from SIS and CUR) were used to address research question 3 (student achievement), Praxis II scores from TCMS were used to address research question 4a (content knowledge among teacher candidates), and course assignment data from the TSA website were used to address research question 4c (teacher assignment in shortage areas; see table A1).

Table A1. Research questions and data sources by key measures

Research question	Outcome measures and data source	Other key variables and data sources
1. Was implementation of Believe and Prepare associated with higher in-service teacher performance ratings?	Teacher in-service performance ratings (continuous outcome, CIS)	• Treatment status for teacher (MRD). • Teacher characteristics (PEP).
2. Was implementation of Believe and Prepare associated with higher teacher retention rates in Louisiana public schools at the school, district, or state level?	Teacher retention in the school, district, or state for one year or three years (binary outcome, PEP)	• Treatment status for teacher (MRD). • Teacher characteristics (PEP).
3. Was implementation of Believe and Prepare in pilot years associated with higher standardized test scores among students with similar prior scores and characteristics?	Student standardized test scores (continuous outcome, SIS)	• Treatment status for student's teacher (MRD). • Teacher characteristics (PEP). • Student-teacher links (CUR). • Student characteristics (SIS).
4a. Did Believe and Prepare teachers have greater competency, as measured by Praxis II scores, than comparison teachers?	Praxis II scores (continuous outcome, TCMS)	• Treatment status for teacher (MRD). • Teacher characteristics (PEP). • Praxis I scores (TCMS).
4b. Did Believe and Prepare teachers teach in the school where they completed their residency more often than comparison teachers?	Teacher taught in the school where the teacher completed a residency (binary outcome, PEP)	• Treatment status for teacher (MRD). • Teacher characteristics (PEP).
4c. Did Believe and Prepare teachers fill teaching positions in shortage areas more often than comparison teachers?	Teacher taught in a shortage area, defined in terms of subject matter or grade (binary outcome, TSA)	• Treatment status for teacher (MRD). • Teacher characteristics (PEP). • Teacher endorsement area (TCMS).
4d. Did Believe and Prepare teachers teach in rural schools more often than comparison teachers?	Teacher taught a rural school (binary outcome, PEP)	• Treatment status for teacher (MRD). • Teacher characteristics (PEP).

CIS is Compass Information System. CUR is Curriculum System. MRD is Mentor and Resident Data. PEP is Profile of Educational Personnel. SIS is Student Information System. TCMS is Teacher Certification Management System. TSA is Teacher Shortage Areas website.
Source: Authors' compilation.

Data preparation

The Louisiana Department of Education provided the following administrative data files with anonymized student, teacher, district, and teacher preparation program IDs. The data included in each file are described below.

- *Teacher.* This file drew data from the PEP collection and includes school year, teacher ID, district ID, sex, race/ethnicity, education level, job assignment (teacher, administrator, or counselor), job function, and experience. The study team used these data to create indicator variables for teacher race/ethnicity (Black, Hispanic, Other race/ethnicity, and White), female teachers, education level (below bachelor's, bachelor's, and master's or higher), and years of experience (with each year of experience as a separate indicator variable). Finally, the study team aggregated job functions into five categories: kindergarten, elementary, secondary, special education, and gifted education.
- *Teacher grade.* This file drew data from the PEP collection and includes school year, teacher ID, school ID, district ID, and the subjects and grades that a teacher taught. The study team combined school and district IDs to identify unique schools and combined subjects into broader subject areas to match those used in the TSA data published by the U.S. Department of Education (more details about TSA data are below). Grades were a series of indicator variables ranging from preK to 12. School locality was coded based on the National Center for Education Statistics' urban-centric school locale assignment system.

  Because some teachers taught in multiple schools during a school year, the study team collapsed the data file to the teacher-school-year level to construct teacher retention variables at the school level. The one-year school-level retention indicator equaled 1 if a teacher taught in the same school in the current year and the next year and 0 otherwise. The three-year school-level retention indicator equaled 1 if a teacher taught in the same school in the current year and each of the three subsequent years. For teachers who taught in multiple schools in a year, the retention indicator equaled 1 for the school or schools where the teacher taught in both the current year and the next (one or three) years. Similarly, because some teachers taught in multiple districts during a school year, teacher retention variables at the district level were constructed using teacher-district-year level data. For teacher retention in the state, data at the teacher-year level were used. Teacher retention variables at the district and state levels were defined in the same way as school-level retention variables.
- *BPFlag.* This file drew data from the MRD collection and contained teacher IDs and an indicator variable for whether a teacher completed a yearlong residency as required under Believe and Prepare. The Louisiana Department of Education paid a stipend to teachers who completed a yearlong residency. This payment database provides accurate tracking of Believe and Prepare teacher residents. The Louisiana Department of Education has an established process for matching records on preservice teachers and teacher residents with Louisiana Department of Education employment records. The process consisted of the following steps (Wan et al., 2021). After the Louisiana Department of Education received lists of residency participants from grantees (teacher preparation institutions or K-12 school systems), its staff attempted to match each participant to a teacher in either the TCMS or the PEP collection. Social Security numbers, when available, were used to match the records. When a Social Security number was not available, a combination of participant information (such as name, teacher preparation program, and K-12 school system) was used.
- *TPP completer roster.* This file was based on the MRD collection and included teacher candidate ID, teacher preparation institution ID, completion year, pathway type, and up to three preservice district and school IDs. The pathway type distinguishes whether an institution is at the undergraduate or graduate level, and

the study team used pathway type in combination with teacher preparation institution IDs to restrict the analytic samples to the 18 traditional, undergraduate teacher preparation institutions. The study team compared preservice district and school IDs with in-service school and district assignments to construct the dependent variable for whether a teacher taught in the school or schools where the teacher completed a residency.

- *Praxis data.* This file drew data from the TCSM and contained teacher ID, teacher preparation institution ID, academic major, pathway type, certification area, test name, test date, test score, and cutscore. About 5.7 percent of these records did not have teacher IDs, likely because those test takers were not later hired by Louisiana public schools. There were 78 test names, some of which had gone through multiple editions over the years. In 13 cases (affecting 3.9 percent of teachers), the test edition changed during a year, and it is reflected in midyear cutscore changes. As a result, the study team normalized test scores by test name, year, and edition (as proxied by the cutscore).

  The study team used test names containing "Praxis I" or, after September 1, 2014, "Core Academic Skills for Educators," to identify Praxis I test scores in math, reading, and writing. The remaining tests were Praxis II tests, and the study team categorized them as special education, content, pedagogy, or other tests. When a teacher had multiple test scores in the same Praxis II test category, the study team averaged those scores to derive a category-specific score. The study team calculated an overall Praxis II score by averaging all Praxis II scores associated with a teacher.

  The study team used certification areas in this data file to identify the programs at which teacher candidates completed their training. This step was necessary because information on which preparation program teachers completed was not available to the study team. More details about how certification areas were used to derive program information is described in the "Identifying teacher preparation programs" section below.

- *Teacher Compass final.* This file drew data from the CIS and included school year, teacher ID, overall performance rating, student growth rating, and professional practice rating. The student growth rating and professional practice rating each contribute 50 percent to the overall performance rating. All ratings were provided on a four-point scale, and the study team normalized them by year to have a mean of 0 and a standard deviation of 1.

- *Teacher shortage area report data.* This file drew data from the TSA website and included the disciplines, subject areas, and grades in which Louisiana reported staffing challenges between 2012 and 2020. The study team merged this file with the Teacher grade file. If any of the subject-grades taught by a teacher in a particular year and school matched those in the teacher shortage area data, that teacher was considered to be teaching in a shortage area in that year and school.

- *Teacher–student linkage.* This file drew data from the CUR and included school year, district ID, school ID, teacher ID, class ID, class period, course name, course type, and course category. The study team linked this file with the Student course file (discussed below) using school year, district ID, school ID, class ID, and class period to identify the students that a teacher taught.

Student data files included:

- *Student course.* This file drew data from the SIS and included school year, district ID, school ID, student ID, course name, course type, and course category. The study team linked this file with the Teacher–student linkage file using school year, district ID, school ID, class ID, and class period to identify the students that a teacher taught.

- *Student information.* This file drew data from the SIS and included school year, district ID, school ID, student ID, sex, race/ethnicity, grade placement, eligibility for the National School Lunch Program, English learner status, and special education participation. The study team used these data to create indicator variables for student race/ethnicity (Black, Hispanic, Other race/ethnicity, and White), female students, whether the student repeated a grade, and whether the student changed schools within a school year.
- *Student test scores.* This file drew data from the SIS and included school year, student ID, grade, and raw scale scores on math and English language arts tests from the 2011/12 through 2018/19 school years. During the study period Louisiana administered several different tests to students in grades 3–8. From 2011/12 through 2013/14, students took the Louisiana Educational Assessment Program (LEAP) and Integrated LEAP test, whereas in 2014/15 students took the Partnership for Assessment of Readiness for College and Careers test. Then from 2015/16 through 2018/19, students took the LEAP 2025 test. The study team used this file to calculate standardized test scores by year and grade, as well as prior year test scores. Because the study team standardized test scores within year and grade, there are limited concerns about switching tests across the study period because each student was measured against peers taking the same test. The study could be affected by switching tests if Believe and Prepare teachers were systematically assigned to certain types of students and the discriminating power of test items changed over time. For example, if Believe and Prepare teachers tended to be assigned to lower-performing students and the discriminating power of tests increased over time, students taught by Believe and Prepare teachers might have become more different from students taught by comparison teachers in later periods because later tests could identify lower-performing students more accurately than earlier tests.

Identifying teacher preparation programs. Information on the specific teacher preparation program that each teacher candidate completed was not available in the administrative records. As a result, the study team used teacher certification areas contained in the TCMS as a proxy for teacher preparation programs. Discussions with staff at the Louisiana Department of Education suggested that this approach was reasonable because of the close correspondence between programs and certification areas. The only uncertainty was that some teachers (14 percent) had multiple certification areas. For these cases the study team treated the combination of multiple certification areas as a unique program. This approach yielded 286 programs over eight years, 83 of which implemented Believe and Prepare during this period. For reference, Title II data show that the 18 traditional undergraduate teacher preparation providers reported a total of 290 programs between 2012/13 and 2018/19, the most recent year for which such data are publicly available. Therefore, certification areas appear to be a fairly accurate approximation for teacher preparation programs.

As a robustness test, the study team used an alternative approach to handle cases in which multiple certification areas were reported. Under this approach teachers with multiple certification areas were considered to have completed multiple programs. This approach yielded 227 programs, 75 of which implemented Believe and Prepare during an eight-year period. The two approaches yielded similar counts.

Verification of Believe and Prepare completion and imputation. The Louisiana Department of Education worked with teacher preparation providers to verify the programs and years in which Believe and Prepare was implemented. Because the Believe and Prepare requirements applied only to teacher candidates newly enrolled in programs, programs implementing Believe and Prepare in theory could have had teacher candidates in the

comparison group.[1] For example, two candidates might have completed the same program in the same year, but one candidate might have entered the program after the Believe and Prepare requirements took effect and would have been subject to them, while the other candidate might have entered the program before the requirements took effect and would not have been subject to them. However, the Louisiana Department of Education indicated that programs were unwilling to run two sets of requirements at the same time in practice and implemented the same requirements for all students in a graduating class, regardless of when a student started. Thus, treatment status was not expected to vary among graduates in the same cohort in the same program. Administrative records compiled by the Louisiana Department of Education were largely consistent with this expectation. Of 1,116 program-years, treated and untreated teacher candidates were both present in only 21 (or 1.9 percent) program-years. Because the treatment was implemented at the program level, the study team assumed that all teachers were subject to Believe and Prepare requirements when more than 50 percent of program graduates were subject to them and that no teacher was subject to Believe and Prepare requirements when 50 percent or less were. This imputation process changed the reported individual treatment status for 0.8 percent (61 out of 7,921) of teacher candidates, with 33 changed from treated to comparison and 28 changed from comparison to treated.

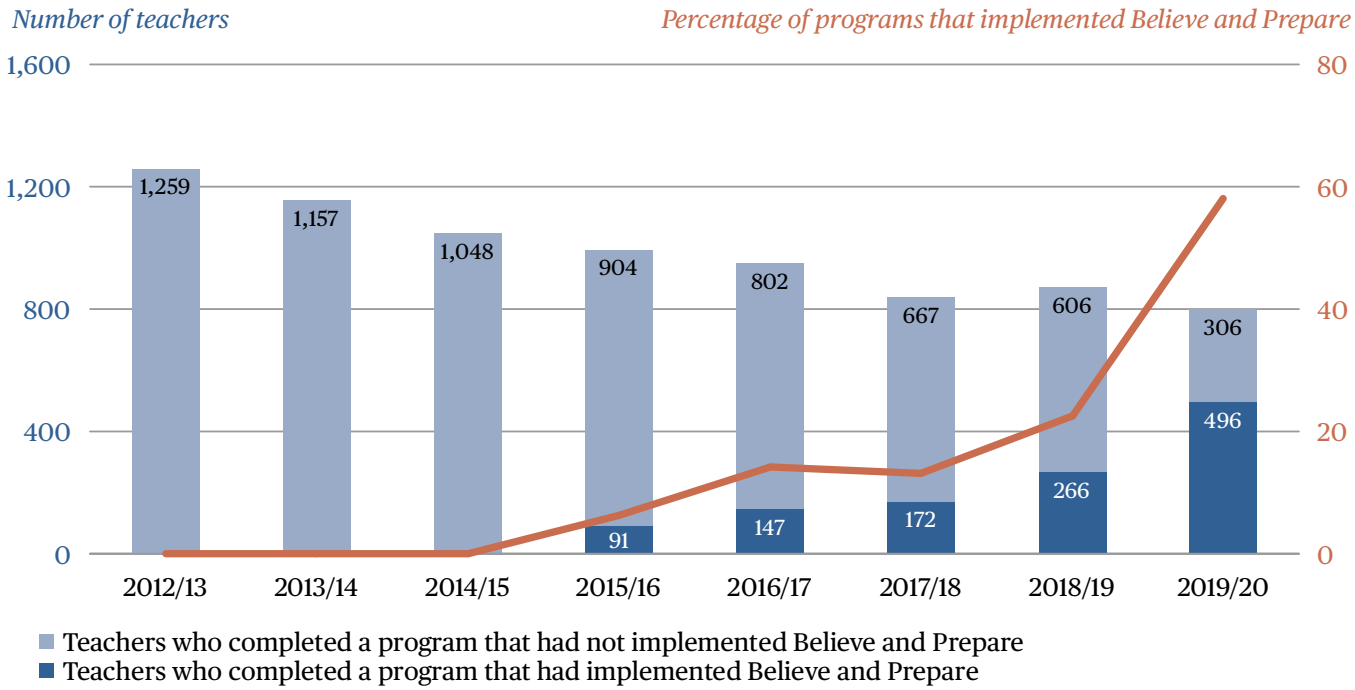
Samples

Although the Believe and Prepare pilot started in 2014/15, teacher preparation programs did not graduate teacher candidates who completed the Believe and Prepare requirements until 2015/16. Since then, the percentage of completers from traditional undergraduate preparation programs that had implemented Believe and Prepare gradually increased to 62 percent in 2019/20 (496 / [496 + 306] = 62 percent; figure A1). By 2019/20, 58 percent of traditional undergraduate teacher preparation programs had graduated teacher candidates who had completed the Believe and Prepare requirements (see figure A1). In 4 of the 18 traditional undergraduate preparation institutions, no programs had produced any graduates who had completed the requirements as of 2019/20.

The increasing share of programs that implemented Believe and Prepare was accompanied by a decline in the total number of teacher candidates in Louisiana completing a traditional undergraduate preparation program. This decline is consistent with the national trend. For example, the American Association of Colleges for Teacher Education (2022) found that the number of people completing a traditional teacher preparation program declined by almost a third between 2008/09 and 2018/19. In addition, the decline in Louisiana started before Believe and Prepare was rolled out and continued almost linearly.

[1] For instance, programs A and B might both have adopted Believe and Prepare in the 2015/16 school year, but if program A admitted teacher candidates as freshmen and program B admitted them as sophomores, then the year teacher candidates were first required to complete a yearlong residency was the 2018/19 school year for program A but 2017/18 for program B (because the candidates were only three years out from their residency when admitted into the program as sophomores).

Figure A1. Percentage of traditional undergraduate teacher preparation programs implementing Believe and Prepare and number of candidates who graduated, by Believe and Prepare implementation status and graduating cohort, 2012/13–2019/20



Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Of the 7,921 teacher candidates who graduated from the 18 traditional undergraduate preparation institutions between 2012/13 and 2019/20, 6,131 were hired as teachers in Louisiana public schools, suggesting an average hiring rate of 77 percent. Naturally, the hiring rates of teacher candidates were higher for older cohorts (about 79 percent), which had more opportunities to be hired than the more recent cohorts (73 percent for the most recent cohort, which represents the rate of hiring immediately after program completion).[2] It is difficult to make simple comparisons of hiring rates between programs that had implemented Believe and Prepare and programs that had not because implementation was staggered, so the comparison set varies depending on when a treated program first implemented Believe and Prepare. For these reasons the study team investigated the possibility of differential hiring rates associated with Believe and Prepare using a two-way fixed effects framework described in the "Sensitivity analysis and robustness checks" section below. That analysis found no statistically significant association between hiring rates and the implementation of Believe and Prepare.

Because administrative records were used for the study, missing values of variables were not a major issue. Among hired teachers who completed a traditional undergraduate preparation program during the study period, regardless of whether it had implemented Believe and Prepare, 98 percent had complete information on demographic characteristics (sex and race/ethnicity), highest degree attained, and job assignment. The study team dropped observations with missing values for covariates in the main analyses. However, as a robustness check, the study team conducted analyses that included all observations with imputation flags added for observations with missing covariate values. The results were nearly identical regardless of whether imputation flags were included for covariates.

[2] This finding is consistent with the 70 percent immediate hiring rate suggested by the Louisiana Department of Education (personal communications, October 22, 2019).

Table A2 compares the baseline characteristics of teachers who completed a preparation program that had implemented Believe and Prepare and of teachers who completed a program that had not implemented it. Tables A3 and A4 compare the baseline characteristics of students whose teachers who completed a teacher preparation program that had implemented Believe and Prepare and of students whose teachers completed a program that had not implemented it. Because implementation timing varied across programs, the years that constitute the baseline years vary by program. The study team therefore constructed summary statistics for treated programs using program-specific baseline years. The years that constitute baseline years for never-treated programs vary depending on the treated programs they are compared with. As a result, the years that should be used to construct baseline characteristics for never-treated programs could not be determined easily. Tables A2–A4 thus include all the years for never-treated programs in summary statistics.

With this caveat in mind, table A2 shows that teachers who completed a preparation program that later implemented Believe and Prepare were less likely than teachers who completed a program that never implemented it to be Hispanic. In addition, tables A3 and A4 show that students whose teachers completed a program that later implemented Believe and Prepare and students whose teachers completed program that never implemented Believe and Prepare were not statistically different on most characteristics. As described in the next section, the study team used a two-way fixed effects framework to account for these baseline differences in teacher and student attributes.

Table A2. Summary of baseline characteristics of teachers who completed a preparation program, by Believe and Prepare implementation status, 2012/13–2018/19

	Teachers who completed a program that later implemented Believe and Prepare		Teachers who completed a program that never implemented Believe and Prepare		
Characteristic	Mean	Standard deviation	Mean	Standard deviation	Hedges' *g* effect size
Female	0.90	0.30	0.83	0.37	0.21
Race/ethnicity					
Black	0.10	0.30	0.12	0.32	-0.06
Hispanic	0.01**	0.08	0.02	0.12	-0.08
Other race/ethnicity[a]	0.01	0.11	0.01	0.12	-0.01
Highest degree earned					
Bachelor's degree	0.98	0.14	0.98	0.13	-0.02
Master's degree or higher	0.02	0.14	0.02	0.13	0.02
Teaching assignment					
Kindergarten	0.06	0.25	0.07	0.25	-0.02
Secondary	0.17	0.38	0.20	0.40	-0.07
Special education	0.07	0.25	0.09	0.28	-0.07
Gifted	0.02	0.12	0.01	0.11	0.02
Number of teachers	2,303		2,698		
Number of programs	83		194		

** Significant at $p < .01$.

Note: Baseline years are the years before a teacher preparation program first implemented Believe and Prepare and vary by program. Significance tests are based on standard errors clustered at the program level.

a. Includes Asian, Native American, Pacific Islander, and multiple race/ethnicities.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table A3. Summary of baseline characteristics of students whose teachers completed a preparation program, analysis sample for math achievement, by Believe and Prepare implementation status, 2012/13–2018/19

Characteristic	Students whose teachers completed a program that later implemented Believe and Prepare		Students whose teachers completed a program that never implemented Believe and Prepare		Hedges' *g* effect size
	Mean	Standard deviation	Mean	Standard deviation	
Female	0.50	0.50	0.49	0.50	0.02
Black	0.42	0.49	0.36	0.48	0.12
Hispanic	0.06	0.23	0.07	0.24	-0.04
Other race/ethnicity[a]	0.05	0.21	0.05	0.22	0.00
Eligible for the National School Lunch Program	0.60	0.49	0.58	0.49	0.04
In special education	0.09	0.29	0.10	0.30	-0.03
English learner student	0.03	0.17	0.02	0.15	0.06
Moved to a new school	0.00	0.04	0.00	0.04	0.00
Repeated grade	0.01	0.07	0.00	0.06	0.15
Number of student-year observations	32,773		38,605		

Note: Baseline years are the years before a teacher preparation program first implemented Believe and Prepare and vary by program. No results were significant at $p < .05$.

a. Includes Asian, Native American, Pacific Islander, and multiple race/ethnicities.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table A4. Summary of baseline characteristics of students whose teachers completed a preparation program, analysis sample for English language arts achievement, by Believe and Prepare implementation status, 2012/13–2018/19

Characteristic	Students whose teachers completed a program that later implemented Believe and Prepare		Students whose teachers completed a program that never implemented Believe and Prepare		Hedges' *g* effect size
	Mean	Standard deviation	Mean	Standard deviation	
Female	0.49	0.50	0.50	0.50	-0.02
Black	0.46**	0.49	0.35	0.48	0.23
Hispanic	0.05	0.22	0.07	0.25	-0.09
Other race/ethnicity[a]	0.05	0.21	0.04	0.21	0.05
Eligible for the National School Lunch Program	0.61**	0.49	0.56	0.50	0.10
In special education	0.09	0.28	0.10	0.30	-0.03
English learner student	0.02	0.15	0.02	0.14	0.00
Moved to a new school	0.00	0.05	0.00	0.04	0.00
Repeated grade	0.00	0.07	0.00	0.06	0.00
Number of student-year observations	62,056		51,196		

** Significant at $p < .01$.

Note: Baseline years are the years before a teacher preparation program first implemented Believe and Prepare and vary by program.

a. Includes Asian, Native American, Pacific Islander, and multiple races/ethnicities.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Data analysis

The timing of Believe and Prepare implementation varied both across traditional undergraduate teacher preparation institutions and across programs within each institution. For example, within the same institution the elementary education program might have implemented Believe and Prepare earlier than the special education program. Treatment assignment (that is, the implementation of Believe and Prepare) occurred at the program level, not at the institution level or the individual teacher candidate level.

A two-way fixed effects (TWFE) framework was used to compare the outcomes of teachers who completed a preparation program that had not implemented Believe and Prepare (comparison group) with teachers who completed a program that had implemented it (treatment group) while persistent program-specific attributes (that is, program fixed effects) and cohort-specific effects applicable to all programs were accounted for. A cohort is defined as a group of teacher candidates who completed a program in the same year. TWFE has been used widely in public policy evaluation (Bertrand et al., 2004). When treatment timing varies across groups, TWFE estimates the weighted average of all possible two-group, two-period difference-in-differences estimates for which the weights are proportional to group sizes and the variance of the treatment variable within each contrast (Goodman-Bacon, 2021).

Research question 1: Was implementation of Believe and Prepare associated with higher in-service teacher performance ratings? The following model with a teacher-level outcome was used to address research question 1:

$$y_{jpct} = \alpha_c + \alpha_p + \beta_1 TREAT_{jpc} + \beta_2 Teacher_{jt} + \sum_k \gamma_k 1\{Exp_{jt} = k\} + \gamma_t + \gamma_d + e_{jpct} \qquad \text{(A1)}$$

where y_{jpct} is the in-service performance rating of teacher *j* in year *t* who received a teaching credential in cohort *c* from program *p*. Both the overall performance rating as well as its equally weighted subcomponents, student growth and professional practice, were examined, and y_{jpct} was examined both as a normalized continuous measure using ordinary least squares and as binary outcomes (rated proficient or higher or rated highly effective) using linear probability models. α_c is teacher cohort fixed effects, and α_p is teacher preparation program fixed effects. $TREAT_{jpc}$ is an indicator variable that equals 1 if a teacher *j* completed a program *p* that had implemented Believe and Prepare in cohort *c*. Details about how this variable was created are in the "Data preparation" section above. $Teacher_{jt}$ includes teacher attributes (sex, race/ethnicity, highest degree attained, and job function). Exp_{jt} is the experience of teacher *j* in year *t*. It does not include the one-year residency. Each year of experience was entered into the equation as a separate indicator variable. To account for potential cross-district and cross-year differences in how teachers were evaluated, equation A1 also controls for district fixed effects, γ_d, and year fixed effects, γ_t. β_1 estimates the Believe and Prepare effect, under the assumption that differences in in-service performance ratings between teachers who completed a program that had implemented Believe and Prepare and teachers who a completed program that had not implemented it would have remained the same had Believe and Prepare not been introduced. Because performance ratings are likely correlated among teachers from the same schools, standard errors were clustered at the school level.

Research question 2: Was implementation of Believe and Prepare associated with higher teacher retention rates in Louisiana public schools at the school, district, or state level? The following model with a teacher-level binary outcome was estimated using logistic regression to address research question 2:

$$P(y_{jpct} = 1) = \alpha_c + \alpha_p + \beta_1 TREAT_{jpc} + \beta_2 Teacher_{jt} + \sum_k \gamma_k 1\{Exp_{jt} = k\} + \gamma_t + \gamma_d + e_{jpct} \qquad \text{(A2)}$$

where y_{jpct} is a retention indicator in year *t* for teacher *j* from program *p* who graduated in cohort *c*. The study examined both one-year retention (whether a teacher in year *t* returned in year $t + 1$) and three-year retention (whether a teacher in year *t* returned in each of the next three consecutive years) at the school, district, and state

levels. The remaining terms have the same definition as in equation A1. Standard errors were clustered at the school level.[3]

Research question 3: Was implementation of Believe and Prepare in pilot years associated with higher standardized test scores among students with similar prior scores and characteristics? The following model with student test score as the outcome was used to address research question 3:

$$Y_{ijpst} = \alpha_c + \alpha_p + \beta_1 TREAT_{jpc} + \beta_2 Y_{i(t-1)} + \beta_3 X_{it} + \beta_4 Teacher_{jt} + \sum_k \gamma_k 1\{Exp_{jt} = k\} + e_{ijpst} \quad \text{(A3)}$$

where Y_{ijpst} is the state test score for each student i with teacher j from program p in subject s (math or English language arts) in year t, normalized within year and grade using state average test scores, and $Y_{i(t-1)}$ is a vector of student i's scores in math and English language arts from the previous year, also normalized within year and grade. The need to control for a student's prior achievement in the same subject is the reason the analysis was restricted to math and English language arts in grades 4–8. X_{it} is a vector of student attributes in year t (sex, race/ethnicity, eligibility for the National School Lunch Program, English learner status, gifted status, special education status, learning disability status), and $Teacher_{jt}$ includes teacher attributes (race/ethnicity, sex, and highest degree attained). α_c, α_p, $TREAT_{jpc}$, and Exp_{jt} have the same definition as in equation A1. β_1 estimates the Believe and Prepare effect, under the assumption that pretreatment differences in student test scores that are attributable to teachers who completed different teacher preparation programs would have remained the same had Believe and Prepare not been introduced. To account for co-teaching of classes, the model was estimated using probability weights, where the weight was defined as 1 / (number of observations per student per year), in a method known as the full roster method (Hock & Isenberg, 2017). Standard errors were clustered at the classroom level.

Research question 4a: Did Believe and Prepare teachers have greater competency, as measured by Praxis II scores, than comparison teachers? The following model with a continuous teacher-level outcome was used to address research question 4a:

$$y_{jpc} = \alpha_c + \alpha_p + \beta_1 TREAT_{jpc} + \beta_2 Teacher_j + \beta_s PraxisI_s + e_{jpc} \quad \text{(A4)}$$

where y is the Praxis II score for teacher j from program p who graduated in cohort c. Praxis II scores were normalized by certification area, year, and test version. These scores were aggregated by type into overall averages, average content test scores, average pedagogy scores, average elementary education scores, average special education scores, and subject-specific scores in math, English language arts, science, and social studies. These aggregate scores were used as the dependent variable. $Teacher_j$ includes teacher sex and race/ethnicity. $PraxisI_s$ is Praxis I scores in subjects s (math, reading, and writing). The scores were normalized by subject, year, and test version and were included to account for variation in ability when a candidate applied to a teacher preparation program. Standard errors were clustered at the teacher preparation program level.

[3] Standard errors were also clustered at the teacher preparation program level in an alternative specification. Standard errors clustered at the school level were more conservative than program-clustered standard errors for school- and district-level retention outcomes but less conservative for state-level retention outcomes. In all cases the estimated standard errors were very similar, and statistical inference was not affected by how standard errors were clustered.

Research question 4b: Did Believe and Prepare teachers teach in the school where they completed their residency more often than comparison teachers?

Research question 4c: Did Believe and Prepare teachers fill teaching positions in shortage areas more often than comparison teachers?

Research question 4d: Did Believe and Prepare teachers teach in rural schools more often than comparison teachers?
The following model with a teacher-level binary outcome was estimated using logistic regression to address research questions 4b, 4c, and 4d:

$$P(y_{jpc} = 1) = \alpha_c + \alpha_p + \beta_1 TREAT_{jpc} + \beta_2 Teacher_j + e_{jpc} \qquad \text{(A5)}$$

where y is a binary variable indicating whether teacher j from program p who graduated in cohort c was placed in the school in which the teacher completed a residency, filled a teaching position in a shortage area, or was placed in a rural school. The study generated measures of whether a teacher ever received a specific type of job assignment during the study period and whether a teacher received a specific type of assignment immediately after completing a teacher preparation program. $Teacher_j$ includes teacher sex and race/ethnicity. Standard errors were clustered at the teacher preparation program level.

Sensitivity analysis and robustness checks

Differential rate of hiring. As discussed in the previous section, 70–80 percent of teacher candidates who completed a traditional undergraduate teacher preparation program were hired to teach in a Louisiana public school. Differential hiring rates associated with Believe and Prepare implementation, if they exist, could suggest a mechanism through which Believe and Prepare might have affected student and teacher outcomes. This question was addressed using a logistic regression model similar to equation A5 but without teacher covariates because these covariates are available only for teachers hired by Louisiana public schools. The dependent variable was whether a teacher was hired immediately after program completion. The results show no statistically significant association between hiring rates and Believe and Prepare implementation. The estimated difference in hiring rate is 0.3 percentage point, with a program-level clustered standard error of 2.1.

Parallel trends. The study's key parameter of interest is the average treatment effect on the treated (ATT), which compares observed outcomes with counterfactual outcomes that could have occurred had programs not implemented Believe and Prepare. Because the counterfactual is not observable, ATT can be identified only using untreated programs under the assumption that the average outcomes for the treated and comparison groups would have followed parallel trends in the absence of treatment. Under this assumption any divergence in postreform outcomes between teachers who completed a preparation program that had implemented Believe and Prepare and teachers who completed a program that had not implemented it can be interpreted as the causal impact of the reform.

One way to test for parallel trends is to estimate a TWFE regression with dynamic treatment effects in the following form (also known as an event study design):

$$P(y_{jpct} = 1) = \alpha_c + \alpha_p + \sum_{l=-K, l\neq -1}^{L} \beta_l 1\{l_p = t - g_p\} + \gamma Teacher_{jt} + \sum_k \gamma_k 1\{Exp_{jt} = k\} + e_{jpct} \qquad \text{(A6)}$$

The study team used research question 2 as an illustration because this is where Believe and Prepare was found to be significantly associated with teacher outcomes. It is therefore important to investigate the extent to which the estimated relationship has a causal interpretation. Compared with equation A2, the single indicator variable for the year in which a program implemented Believe and Prepare ($TREAT_{jpc}$) was replaced with a series of relative time binary variables, $1\{l_p = t - g_p\}$, that equal 1 if program p first implemented Believe and Prepare in

year g_p. Thus, $l = 0$ for the year in which Believe and Prepare was first implemented. As is conventional, the year immediately before the implementation ($l = -1$) is the reference period and is omitted. β_l estimates the cumulative effect $l + 1$ periods after the inception of Believe and Prepare for $l \geq 0$. When $l < -1$, β_l estimates the placebo effect $|l|$ periods before Believe and Prepare took effect, and these estimates provide a test for the parallel trends assumption.

This model cannot accommodate the 4.8 percent of programs that switched in and out of Believe and Prepare. Those programs were dropped for this analysis. In addition, to facilitate comparisons with estimators that are robust to staggered implementation and heterogeneous treatment effects (discussed next), equation A6 was estimated as a linear probability model rather than a logistic regression as in the main analysis.[4] These changes had minimal impact on the estimated average treatment effect. Compared with the logistic regression coefficient from the main analysis (reproduced in column 1 in table C1 in appendix C as a marginal effect), a linear probability model (column 2) produced an identical estimate up to two decimal places. The estimated effect based on a sample that excluded programs that switched in and out is 0.03 (see column 3 in table C1), which is similar to the marginal effect of 0.02 that was reported in the main analysis.

Estimates of β_l are reported in panel A of figure C1 in appendix C. None of the estimates for periods before Believe and Prepare took effect is statistically significant. The joint parallel pretrends test is also not statistically significant ($p = .86$; see column 4 in table C1), suggesting that the parallel trends assumption was not violated. For the post-treatment period, the coefficient for period $t + 1$ (two years after initial implementation of Believe and Prepare) is statistically significant ($p = .04$).

Staggered implementation and heterogeneous effects. In a traditional "static" TWFE model, as specified in equation A3, the estimated overall treatment effect would be biased when program implementation timing is staggered and treatment effects vary across group and time periods (de Chaisemartin & D'Haultfœuille, 2020; Goodman-Bacon, 2021). This is because the TWFE estimator is the weighted average of all two-group, two-period comparisons in the data, in which the weights are proportional to group sizes and the variance of the treatment variable within each contrast (Goodman-Bacon, 2021). de Chaisemartin and D'Haultfœuille (2020) showed that the weights are positively related to the number of observations in a particular group and time period (g, t) and the average number of treatments across groups and periods and negatively related to the average sample size across periods for group g and the average sample size across groups at period t. As such, these weights have no direct policy relevance, rendering the estimated average treatment effect uninterpretable. In cases where already treated groups are used as the comparison for groups that have not yet implemented Believe and Prepare, some of the weights can be negative. Furthermore, Sun and Abraham (2021) showed that in dynamic TWFE models such as equation A6, each β_l is contaminated by both the weights and effects from periods $l' \neq l$. Therefore, Sun and Abraham concluded that unless treatment effects are homogeneous, β_l for $l < -1$ cannot be used to test parallel trends despite their wide use in existing applied research.

Several empirical strategies have been developed that are robust to heterogenous treatment effects (for example, Borusyak et al., 2021; Callaway & Sant'Anna, 2021; de Chaisemartin & D'Haultfœuille, 2020; Sun & Abraham, 2021). These estimators differ in terms of the type of data they handle (panel versus repeated cross-sections), whether they allow for dynamic treatment effects, parallel trends assumptions (conditional versus unconditional), whether the treatment is binary or nonbinary, and whether treatment assignment is staggered (that is, irreversible treatment assignment versus allowing for switch in-and-out). What is common across these

[4] Although linear probability is a reasonable approximation (personal communications with de Chaisemartin), Woodridge (2021) suggests that more research is needed for when the outcome is binary.

new estimators is that they carefully choose comparison sets and weights that have policy relevance. Because the framework proposed by Callaway and Sant'Anna (2021) appears to be more general and to require less stringent assumptions,[5] the current study estimated their doubly robust estimator based on stabilized inverse probability weighting and ordinary least squares (DRIPW) to assess how sensitive the main findings are to staggered implementation and heterogeneous treatment effects. Because a large number of teacher preparation programs never implemented Believe and Prepare during the study period, the study first used never-treated programs as the comparison group. However, never-treated programs could be inherently different from programs that received treatment at some point in ways that are unobservable. As a result, following Callaway and Sant'Anna's suggestion, the current study also estimated an alternative specification that dropped never-treated programs and used not-yet-treated programs as the comparison group. The tradeoff of this choice is that the sample size was reduced to one-seventh of the sample size when never-treated programs were used as the comparison group.

The DRIPW estimator assumes staggered policy adoption, meaning that once a unit is treated, it remains treated in the following periods. It also assumes limited or no treatment anticipation, which may lead to changes in outcomes before a treatment has been implemented. It also assumes parallel trends (conditional on baseline covariates that could be associated with post-treatment outcome trends in the absence of treatment) between the treated group and either a never-treated comparison group or not-yet-treated comparison group. Finally, it assumes that both treated and comparison units are present along the distribution of the estimated propensity score of treatment participation.

Callaway and Sant'Anna (2021) start by estimating disaggregated two-group, two-period causal parameters (that is, group-time average treatment effect):

$$ATT(g,t) = \mathrm{E}\left[Y_t^1 - Y_t^0 | G_g = 1\right]$$

where $ATT(g, t)$ is the average treatment effect in period t for programs that first implemented Believe and Prepare in period g, t is time period (1, …, T), g is time when a treatment first took effect, G_g is a binary variable that equals 1 if a program first implemented Believe and Prepare in period g, Y_t^1 is the outcome in period t when a program implemented Believe and Prepare, and Y_t^0 is the outcome in period t when a program did not implement it. $ATT(g, t)$ can be identified using approaches that are based on outcome regressions (Heckman et al., 1997, 1998), inverse probability weighting (Abadie, 2005), or both (Callaway & Sant'Anna, 2021). These approaches differ in how selection into treatment is modeled. The outcome regression-based approach relies on the evolution of outcomes, the inverse probability weighting approach relies on the propensity score of implementing treatment conditional on X, and the DRIPW estimator relies on both. Importantly, the DRIPW estimator is unbiased and consistent as long as either the outcome trends or the propensity score model are correctly specified. Therefore, it is more robust to misspecifications (Callaway & Sant'Anna, 2021).

[5] For example, Sun and Abraham (2021) are a special case of the framework proposed by Callaway and Sant'Anna (2021). On the other hand, de Chaisemartin and D'Haultfœuille (2020) are more general than Callaway and Sant'Anna in that they allow for treated programs to revert to being untreated and treatment can be nonbinary.

For each group and calendar time period (g, t) with never-treated units as the comparison group, the DRIPW estimator is (Callaway & Sant'Anna, 2021, p. 206):

$$\widehat{ATT(g,t)} = \mathrm{E}\left[\left(\frac{G_g}{\mathrm{E}[G_g]} - \frac{\frac{\widehat{p_g(X)}C}{1-\widehat{p_g(X)}}}{\mathrm{E}\left[\frac{\widehat{p_g(X)}C}{1-\widehat{p_g(X)}}\right]}\right)\left(Y_t - Y_{g-1} - \widehat{m_{g,t}^{never}}(X)\right)\right]$$

where t is time period (1, ..., T); g is time when a treatment first took effect; G_g is a binary variable that equals 1 if a program first implemented Believe and Prepare in period g; C is a binary variable that equals 1 for programs that never implemented Believe and Prepare (when G_g equals 1, C equals 0, and vice versa); X is baseline covariates, which include demographic characteristics, education level, years of teaching experience, and job functions, aggregated to the program level for periods before Believe and Prepare was first implemented; $\widehat{m_{g,t}^{never}}$ is the estimated difference in outcome before and after treatment for the never-treated group (an estimate of $\mathrm{E}[Y_t - Y_{g-1}|X, C = 1]$); and $\widehat{p_g(X)}$ is the estimated generalized propensity score for implementing treatment in period g conditional on X.

In the next step the estimated group-time average treatment effects can be aggregated into summary causal estimates that are of policy interest. To test for the parallel trends assumption, for example, the estimated group-time effects can be aggregated by elapsed treatment time using the weight (Callaway & Sant'Anna, 2021):

$$w(g,t) = 1\{g + l \leq T\}1\{t - g = l\}P(G = g|G + l \leq T)$$

where t is time period (1, ..., T), G is time when a treatment first took effect, and l is time elapsed since Believe and Prepare was first implemented. The estimated treatment effects aggregated by elapsed time can then be compared with the dynamic TWFE estimates produced by equation A6. These DRIPW estimates are presented in panel B in figure C1 in appendix C (and in column 5 in table C1), with never-treated programs as the comparison group, and in panel C (and in column 6 in table C1), with not-yet-treated programs as the comparison group. In both cases inference follows Callaway and Sant'Anna (2021) and is based on a cluster-robust multiplicative wild bootstrap procedure that accounts for the dependency across different group-time average treatment effect estimators (in other words, the potential multiple-testing problems).

Largely consistent with estimates presented in panel A, the estimates of β_l for $l < -1$ when never-treated programs were used as the comparison group suggest no violation of the parallel pretrend assumption for up to five years before Believe and Prepare was first implemented (see panel B in figure C1 in appendix C). The joint test suggests no violation of the parallel pretrend assumption (p = .29). Using not-yet-treated programs as the comparison group reduced the sample size substantially, resulting in more volatile point estimates and wider confidence intervals (see panel C in figure C1). The results suggest that teacher retention rates four and six years before Believe and Prepare was first implemented were marginally significantly different (p < .10), but the joint test of parallel trends remains statistically nonsignificant (p = .56).

Did the effect of Believe and Prepare vary with how long the reform had been implemented? In addition to providing a robustness check of the parallel trends assumption, there is substantive interest in how the effectiveness of Believe and Prepare might have changed depending on how long it had been implemented by a teacher preparation program. Panel C in figure C1 in appendix C, for example, suggests that the Believe and Prepare effect increased over time. Such comparisons, however, are likely contaminated by changes in the composition of programs. For example, although the effect at the end of the first year (t) can be estimated for all teacher preparation programs that ever implemented Believe and Prepare during the study period, only the earliest

implementers would be used in estimating the $t + 3$ effect, and only the latest implementers would be used in estimating the $t - 6$ effect.

To remove contamination due to compositional changes, the analytic sample was restricted to teachers who completed a preparation program that had at least three pretreatment years and at least three post-treatment years, thus reporting the effect for $l \in [t - 3, t + 2]$. The results are reported in panel D in figure C1 in appendix C. With a stable sample of programs, there seems to be no significant treatment effects for any of the post-treatment periods. The coefficient for period $t - 2$ is statistically significant, but the joint test of parallel trends remains nonsignificant ($p = .20$; see table C1).

Did the Believe and Prepare effect vary depending on when a program implemented the reform? To answer this question, group-time average treatment effects need to be aggregated by g, the period in which a program first implemented Believe and Prepare using weight (Callaway & Sant'Anna, 2021):

$$w(\tilde{g}) = 1\{t \geq g\}1\{g = \tilde{g}\}/(T - g + 1)$$

where a program first implemented Believe and Prepare in period $\tilde{g}$. The results are reported in column 8 of table C1 in appendix C. Teachers who completed a preparation program that had implemented Believe and Prepare early (that is, in 2015, 2016, and 2017) were not statistically different from teachers who completed a program that had not implemented Believe and Prepare in terms of retention rate in Louisiana. However, teachers who completed a program that had implemented Believe and Prepare in 2018 were 22 percentage points more likely than comparison teachers to stay in Louisiana. This could be explained by changes in the ways in which Believe and Prepare was implemented or by a correlation between the timing of program participation and programs' willingness and preparedness for the reform.

Limitations

The current study has four main limitations. First, because of a lack of student test scores in recent years due to the COVID-19 pandemic, the analysis of the association between student achievement and the reform is limited to teachers who completed a preparation program during the first three years of implementation. As a result, the student achievement findings rely on a small sample of teachers who completed a preparation program that had implemented Believe and Prepare. Relatedly, the study was unable to examine other student outcomes that teachers influence, such as absence from school and misbehavior resulting in suspension (for example, see Jackson, 2018), that could be associated with teachers having completed the yearlong residency.

Second, because data on resident-mentor linkage were not collected until the 2019/20 school year, the study could not investigate how mentor quality may be associated with student and teacher (mentee) outcomes or the extent to which student outcomes in mentors' classrooms were affected by hosting student teachers for extended periods. Emerging research suggests that mentor quality is associated with the in-service performance of the teacher candidates they mentor (Goldhaber, Krieg, & Theobald, 2020). States and districts have some control over which teachers serve as mentors, and there is substantial scope for change in mentor assignments (Goldhaber, Krieg, Naito, & Theobald, 2020).

Third, due to a lack of data on teacher preparation program applicants, the study could not investigate the extent to which Believe and Prepare may have altered the number and composition of candidates who apply to and persist in teacher preparation programs. The added costs in time and foregone earnings to teacher candidates due to the shift from a six-week to a yearlong residency might have dissuaded some students from becoming teachers.

Finally, to attribute any observed difference in outcomes between teachers who completed a preparation program that had implemented Believe and Prepare and teachers who completed a program that did not implement it, the study team assumed that those outcomes would have followed parallel trends over time in the absence of reform. Further, the study team assumed that prereform outcomes were not affected by anticipation of the upcoming reform and that no other contemporaneous policy changes affected the outcomes. Even when these assumptions are found to hold true, the estimated effect of Believe and Prepare could be biased if the effect varied over time or across teacher preparation programs. The study investigated the plausibility of these assumptions and the robustness of the main findings to these potential sources of bias, as described in the previous section. Although the findings from these analyses do not contradict the main findings, they have large margins of error. As a result, findings from this study should be interpreted as descriptive rather than causal.

References

Abadie, A. (2005). Semiparametric difference-in-difference estimators. *Review of Economic Studies, 72*(1), 1-19.

American Association of Colleges for Teacher Education. (2022, March 22). *AACTE's national portrait sounds the alarm on declining interest in education careers* [Press Release]. Retrieved October 24, 2022, from https://aacte.org/2022/03/aactes-national-portrait-sounds-the-alarm-on-declining-interest-in-education-careers/.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics, 119*(1), 249-275.

Borusyak, K., Jaravel, X., & Spiess, J. (2021). *Revisiting event study designs: Robust and efficient estimation.* https://doi.org/10.48550/arXiv.2108.12419.

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

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review, 110*(9), 2964-2696.

Goldhaber, D., Krieg, J., Naito, N., & Theobald, R. (2020). Making the most of student teaching: The importance of mentors and scope for change. *Education Finance and Policy, 15*(3), 581-591. https://eric.ed.gov/?id=EJ1259717.

Goldhaber, D., Krieg, J., & Theobald, R. (2020). Effective like me? Does having a more productive mentor improve the productivity of mentees? *Labour Economics, 63*(1), 1-13. https://eric.ed.gov/?id=ED618755.

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

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica, 66*(5), 1017-1098.

Heckman, J. J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605-654.

Hock, H., & Isenberg, E. (2017). Methods for accounting for co-teaching in value-added models. *Statistics and Public Policy, 4*(1), 1-11.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non-test score outcomes. *Journal of Political Economy, 126*(5), 2072-2107. http://dx.doi.org/10.1086/699018.

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

Wan, Y., Nguyen, T., Lazarev, V., Zacamy, J., & Gerdeman, D. (2021). *Outcomes for early career teachers prepared through a pilot residency program in Louisiana* (REL 2021–079). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. https://eric.ed.gov/?id=ED612482.

Wooldridge, J. M. (2021). *Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators*. Department of Economics, Michigan State University. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3906345.

Appendix B. Supporting analyses

This appendix provides supporting analyses for the findings in the report.

Table B1 summarizes the number of teachers who completed their training in one of the 18 traditional undergraduate teacher preparation institutions, by institution and year.

Table B1. Number of preparation program completers in each Louisiana undergraduate teacher preparation institution, by year, 2012/13–2019/20

Institution	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19	2019/20
1	18	12	16	21	9	2	6	4
2	101	106	72	63	64	89	78	51
3	19	23	14	9	5	16	7	9
4	47	35	41	41	33	29	25	22
5	18	23	21	17	19	9	13	16
6	10	13	18	8	12	9	6	7
7	79	73	81	81	80	71	86	53
8	240	218	202	173	195	140	138	124
9	0	0	1	6	6	5	5	12
10	41	68	48	39	40	34	30	28
11	116	101	81	84	87	60	94	74
12	37	6	51	51	67	59	53	63
13	69	41	47	57	38	40	53	52
14	232	187	167	152	158	144	128	129
15	19	14	8	5	11	6	12	10
16	199	222	171	180	120	121	134	145
17	12	11	5	4	5	3	3	2
18	2	4	4	4	0	2	1	1

Note: Institution names are suppressed per the data-sharing agreement between Regional Educational Laboratory Southwest and the Louisiana Department of Education.
Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table B2 reports summary statistics of teacher outcomes, by Believe and Prepare implementation status and year. For each outcome the mean, standard deviation, and number of observations are reported.

Table B2. Teacher outcomes (mean, standard deviation, and number of observations) for teachers who completed a preparation program that had implemented Believe and Prepare and for teachers who completed a program that had not implemented it, by year, 2015/16–2019/20

	2015/16		2016/17		2017/18		2018/19		2019/20	
Outcome and statistic	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers
Teacher performance rating: Overall										
Mean	3.14	3.26	3.21	3.28	3.04	3.21	3.13	3.24	3.34	3.42
Standard deviation	0.42	0.52	0.50	0.52	0.57	0.53	0.56	0.53	0.52	0.48
Number of observations	10	2,220	73	2,752	173	3,182	323	3,491	761	3,630

Outcome and statistic	2015/16		2016/17		2017/18		2018/19		2019/20	
	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers
Teacher performance rating: Student growth										
Mean	3.30	3.31	3.23	3.30	2.95	3.14	3.04	3.14	3.36	3.39
Standard deviation	0.48	0.75	0.71	0.76	0.89	0.81	0.91	0.83	0.77	0.72
Number of observations	10	2,220	73	2,752	173	3,182	323	3,491	761	3,630
Teacher performance rating: Professional practice										
Mean	2.97	3.20	3.19	3.26	3.13	3.29	3.22	3.35	3.32	3.45
Standard deviation	0.46	0.50	0.45	0.49	0.48	0.48	0.47	0.46	0.48	0.44
Number of observations	10	2,220	73	2,752	173	3,182	323	3,491	761	3,630
Retention: One year in the same school										
Mean	–	0.80	0.75	0.79	0.78	0.77	0.81	0.79	0.81	0.81
Standard deviation	–	0.40	0.44	0.41	0.42	0.42	0.39	0.41	0.39	0.39
Number of observations	0	2,322	64	2,924	175	3,467	313	3,797	528	4,100
Retention: One year in the same district										
Mean	–	0.87	0.86	0.85	0.87	0.84	0.89	0.86	0.88	0.87
Standard deviation	–	0.34	0.35	0.36	0.33	0.36	0.32	0.34	0.32	0.33
Number of observations	0	2,242	64	2,804	175	3,284	305	3,592	515	3,860
Retention: One year in the state										
Mean	–	0.93	0.95	0.93	0.95	0.92	0.94	0.92	0.94	0.93
Standard deviation	–	0.25	0.21	0.26	0.22	0.27	0.24	0.27	0.24	0.26
Number of observations	0	2,241	64	2,803	175	3,284	305	3,592	515	3,860
Retention: Three years in the same school										
Mean	–	0.51	0.53	0.50	0.54	0.52	–	–	–	–
Standard deviation	–	0.50	0.50	0.50	0.50	0.50	–	–	–	–
Number of observations	0	2,322	64	2,924	175	3,467	0	0	0	0
Retention: Three years in the same district										
Mean	–	0.65	0.70	0.64	0.72	0.66	–	–	–	–
Standard deviation	–	0.48	0.46	0.48	0.45	0.47	–	–	–	–
Number of observations	0	2,242	64	2,804	175	3,284	0	0	0	0
Retention: Three years in the state										
Mean	–	0.80	0.83	0.79	0.86	0.79	–	–	–	–
Standard deviation	–	0.40	0.38	0.41	0.35	0.41	–	–	–	–
Number of observations	0	2,241	64	2,803	175	3,284	0	0	0	0
Praxis II score: Average										
Mean	-0.05	-0.07	-0.03	-0.09	-0.23	-0.11	-0.08	-0.16	-0.05	-0.25
Standard deviation	0.75	0.77	0.71	0.76	0.73	0.69	0.70	0.68	0.69	0.63
Number of observations	71	692	126	608	142	493	233	443	344	216
Praxis II score: Content										
Mean	-0.13	-0.03	-0.10	-0.04	-0.20	-0.06	-0.06	-0.10	-0.09	-0.17
Standard deviation	0.82	0.84	0.86	0.79	0.69	0.74	0.72	0.72	0.72	0.64
Number of observations	71	559	124	499	128	434	228	372	302	191

Outcome and statistic	2015/16		2016/17		2017/18		2018/19		2019/20	
	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers	B&P teachers	Non-B&P teachers
Praxis II score: Pedagogy										
Mean	0.04	-0.12	0.05	-0.13	-0.28	-0.12	-0.16	-0.23	-0.05	-0.31
Standard deviation	0.93	0.95	0.86	0.97	1.03	0.94	0.97	0.93	0.93	0.86
Number of observations	71	689	125	605	141	490	233	440	344	213
Praxis II score: English										
Mean	-0.04	0.17	-0.36	0.01	-0.28	-0.13	-0.06	-0.12	-0.05	-0.19
Standard deviation	1.07	0.95	0.90	0.91	0.86	0.99	0.97	0.94	0.92	0.83
Number of observations	17	65	23	132	37	148	166	273	257	146
Praxis II score: Math										
Mean	-0.50	0.08	0.08	0.12	-0.06	-0.18	-0.07	-0.07	-0.04	0.02
Standard deviation	–	0.83	1.20	0.90	0.88	0.92	0.95	0.99	0.94	1.01
Number of observations	1	51	16	117	26	135	155	258	242	139
Praxis II score: Science										
Mean	-0.43	-0.23	-0.50	0.12	-0.91	-0.19	0.00	-0.14	-0.11	-0.11
Standard deviation	–	0.71	0.98	0.83	0.90	0.96	0.98	0.90	0.94	0.88
Number of observations	1	37	11	108	23	135	153	245	242	142
Praxis II score: Social studies										
Mean	-0.28	-0.20	-0.33	-0.10	-0.33	-0.32	-0.12	-0.22	-0.17	-0.16
Standard deviation	0.46	0.73	0.82	0.95	1.05	0.78	0.85	0.80	0.84	0.88
Number of observations	16	63	19	121	30	148	172	256	254	156
Praxis II score: Elementary										
Mean	-0.10	-0.03	-0.10	-0.06	-0.27	-0.04	-0.30	-0.20	-0.46	-1.28
Standard deviation	0.83	0.87	0.89	0.82	0.80	0.79	1.10	0.88	1.52	0.94
Number of observations	36	413	89	309	88	237	48	53	10	9
Praxis II score: Special education										
Mean	–	-0.08	–	0.00	–	-0.15	0.00	-0.16	0.81	-0.40
Standard deviation	–	0.91	–	0.95	–	0.94	0.77	0.92	0.72	0.72
Number of observations	0	49	0	62	0	52	9	59	14	38
Placed in the school in which the teacher completed a residency										
Mean	–	0.25	0.21	0.22	0.20	0.20	0.21	0.18	0.21	0.18
Standard deviation	–	0.50	0.41	0.42	0.40	0.40	0.41	0.39	0.41	0.38
Number of observations	0	4	66	603	174	1164	312	1587	526	1966
Placed in a shortage area										
Mean	–	0.92	0.98	0.90	0.97	0.88	0.26	0.25	0.45	0.40
Standard deviation	–	0.27	0.12	0.30	0.18	0.32	0.44	0.43	0.50	0.49
Number of observations	0	2,331	66	2,933	175	3,481	314	3,803	528	4,112
Placed in a rural school										
Mean	–	0.28	0.34	0.28	0.26	0.28	0.30	0.29	0.33	0.29
Standard deviation	–	0.45	0.48	0.45	0.44	0.45	0.46	0.45	0.47	0.46
Number of observations	0	2,300	64	2,891	174	3,412	309	3,718	504	3,991

– is not available due to insufficient data.

B&P teachers are teachers who completed a preparation program that had implemented Believe and Prepare.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table B3 reports summary statistics of test scores for students whose teachers completed a preparation program that had implemented Believe and Prepare and for students whose teachers completed a program that had not implemented it, by year. For each subject, test scores were standardized by year. The mean, standard deviation, and number of observations are reported for each subject and year. The average scores for Believe and Prepare teachers were substantially different from 0 in 2016/17. For example, the average math score for students of Believe and Prepare teachers was 0.33 standard deviation. This estimate was calculated using students of 5 Believe and Prepare teachers, and these teachers were assigned to high-performing classes that had an average prior math score of 0.31 standard deviation.

Table B3. Student outcomes (mean, standard deviation, and number of observations) for students whose teachers completed a preparation program that had implemented Believe and Prepare and for students whose teachers completed a program that had not implemented it, by year, 2016/17–2018/19

	2016/17		2017/18		2018/19	
Outcome and statistic	Students of B&P teachers	Students of non-B&P teachers	Students of B&P teachers	Students of non-B&P teachers	Students of B&P teachers	Students of non-B&P teachers
Standardized math score						
Mean	0.33	0.03	-0.05	0.08	-0.06	0.09
Standard deviation	1.09	0.93	1.01	0.94	0.93	0.96
Number of observations	428	14,059	875	16,742	1,455	17,752
Standardized English language arts score						
Mean	-0.24	0.09	-0.07	0.11	-0.09	0.12
Standard deviation	0.93	0.96	0.85	0.96	0.90	0.97
Number of observations	838	22,026	1,393	25,545	2,776	26,317

B&P teachers are teachers who completed a preparation program that had implemented Believe and Prepare.
Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table B4 is a supplement to figure 1 in the main report. The columns are teacher outcomes, and the rows are key independent variables. The treatment row reports the estimated association between Believe and Prepare and teacher performance ratings. Results from the first three columns are reported in figure 1 in the main report. The last six columns report results for whether a teacher was rated proficient or higher and whether a teacher was rated highly effective. These were estimated as linear probability models because logistic regressions were unable to converge.

Table B4. Regression coefficient and standard error estimates for analyses of the relationship between in-service performance ratings and whether a teacher completed a preparation program that had implemented Believe and Prepare, 2015/16–2019/20

	Average rating			Rated proficient or higher			Rated highly effective		
Variable	Overall	Student growth	Professional practice	Overall	Student growth	Professional practice	Overall	Student growth	Professional practice
Treatment	0.05	0.06	0.02	0.01	0.01	0.00	0.01	0.05**	-0.00
	(0.04)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)
Black	-0.13**	-0.07*	-0.17**	-0.05**	-0.02	-0.06**	-0.01*	-0.03	-0.01
	(0.04)	(0.04)	(0.04)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
Hispanic	-0.06	-0.06	-0.03	-0.04	-0.01	0.01	-0.02*	-0.04	-0.02
	(0.10)	(0.10)	(0.11)	(0.05)	(0.04)	(0.04)	(0.01)	(0.04)	(0.02)
Other race/ethnicity	0.06	0.04	0.07	0.01	0.00	0.04	-0.00	-0.04	-0.01
	(0.08)	(0.08)	(0.07)	(0.04)	(0.03)	(0.03)	(0.01)	(0.03)	(0.02)
Female	0.21**	0.14**	0.24**	0.08**	0.06**	0.09**	0.01*	0.03*	0.02†
	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Bachelor's degree	-0.87	-1.09	-0.21	-0.32	-0.28	0.13	-0.09**	0.13	-0.20**
	(0.92)	(1.17)	(0.17)	(0.23)	(0.28)	(0.13)	(0.03)	(0.29)	(0.05)
Master's degree or higher	-0.79	-1.08	-0.06	-0.30	-0.27	0.15	-0.08*	0.12	-0.17**
	(0.92)	(1.17)	(0.17)	(0.23)	(0.28)	(0.13)	(0.03)	(0.29)	(0.05)
Job function: Kindergarten	0.13**	0.23**	-0.07*	0.06**	0.09**	-0.03*	-0.02*	0.08**	-0.02
	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
Job function: Secondary	-0.01	0.03	-0.07*	0.01	0.04**	-0.01	-0.00	-0.00	-0.01
	(0.04)	(0.03)	(0.04)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
Job function: Special education	-0.01	0.05	-0.09**	0.01	0.03*	-0.01	-0.01	0.03†	-0.01
	(0.04)	(0.04)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Job function: Gifted education	0.24**	0.18**	0.23**	0.08**	0.07**	0.08**	-0.00	0.08**	0.03
	(0.06)	(0.06)	(0.06)	(0.03)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)

Variable	Average rating			Rated proficient or higher			Rated highly effective		
	Overall	Student growth	Professional practice	Overall	Student growth	Professional practice	Overall	Student growth	Professional practice
2 years of experience	0.18**	0.07**	0.28**	0.07**	0.04**	0.10**	0.01**	0.00	0.02**
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)
3 years of experience	0.32**	0.17**	0.42**	0.12**	0.06**	0.16**	0.02**	0.03†	0.03**
	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
4 years of experience	0.38**	0.19**	0.50**	0.14**	0.08**	0.19**	0.02*	0.02	0.05**
	(0.04)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)
5 years of experience	0.38**	0.16**	0.56**	0.15**	0.07**	0.20**	0.03**	0.03	0.07**
	(0.05)	(0.05)	(0.05)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)
6 years of experience	0.43**	0.20**	0.59**	0.18**	0.13**	0.21**	0.02	-0.00	0.06**
	(0.06)	(0.06)	(0.06)	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	(0.02)
7 years of experience	0.40**	0.21**	0.53**	0.15**	0.09**	0.19**	0.05**	0.02	0.09**
	(0.07)	(0.08)	(0.07)	(0.03)	(0.03)	(0.03)	(0.02)	(0.04)	(0.02)
8 years of experience	0.45**	0.25**	0.56**	0.16**	0.11**	0.19**	0.03	0.05	0.06*
	(0.09)	(0.09)	(0.09)	(0.04)	(0.04)	(0.04)	(0.02)	(0.04)	(0.03)
9 years of experience	0.39**	0.17	0.57**	0.14**	0.08†	0.24**	0.02	-0.04	0.07†
	(0.11)	(0.11)	(0.11)	(0.05)	(0.04)	(0.04)	(0.03)	(0.06)	(0.04)
10 years of experience	0.74**	0.52	0.77**	0.22**	0.16†	0.26**	-0.11**	0.09	0.19
	(0.22)	(0.40)	(0.24)	(0.07)	(0.09)	(0.06)	(0.03)	(0.36)	(0.27)
Number of teacher-year observations	19,477	19,477	19,477	19,477	19,477	19,477	19,477	19,477	19,477
R^2	0.24	0.16	0.27	0.17	0.13	0.21	0.09	0.13	0.13

† Significant at $p < .10$; * significant at $p < .05$; ** significant at $p < .01$.

Note: Numbers in parentheses are clustered standard errors at the school level. The model includes indicator variables for teacher experience, as well as cohort, program, district, and year fixed effects.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table B5 is a supplement to figure 2 in the main report. The columns are teacher retention outcomes, and the rows are independent variables. Each model was estimated by a logistic regression, and coefficients are reported as odds ratios. An estimated odds ratio larger than 1 suggests that the treated group was more likely than the comparison group to be observed to have an outcome, whereas an estimate smaller than 1 suggests that the treated group was less likely than the comparison group to do so. To facilitate interpretation, the estimated odds ratios were converted to marginal effects and plotted as figure 2 in the main report. The marginal effect can be interpreted as the difference in the probability of observing an outcome among teachers who completed a preparation program that had implemented Believe and Prepare relative to teachers who completed a program that had not implemented it.

Table B5. Regression coefficient (odds ratios) and standard error estimates for analyses of the relationship between teacher retention and whether a teacher completed a preparation program that had implemented Believe and Prepare, 2016/17–2019/20

	One-year retention			Three-year retention		
Variable	School	District	State	School	District	State
Treatment	1.11	1.23†	1.38*	1.40†	1.46*	1.38
	(0.11)	(0.14)	(0.22)	(0.24)	(0.27)	(0.37)
Black	0.79**	0.90	0.85	0.69**	0.88	0.85
	(0.06)	(0.08)	(0.09)	(0.08)	(0.11)	(0.12)
Hispanic	1.50†	1.42	1.90	1.52	1.30	1.18
	(0.32)	(0.40)	(0.76)	(0.49)	(0.46)	(0.50)
Other race/ethnicity	0.86	0.67*	0.62*	0.76	0.54*	0.45**
	(0.14)	(0.12)	(0.13)	(0.20)	(0.16)	(0.11)
Female	1.16*	1.18*	1.38**	1.14	1.26*	1.48**
	(0.08)	(0.10)	(0.15)	(0.11)	(0.14)	(0.19)
Bachelor's degree	1.92	2.33	18.69*	0.69	0.48	0.67
	(1.82)	(3.43)	(22.08)	(0.47)	(0.60)	(0.60)
Master's degree or higher	1.36	1.71	12.74*	0.46	0.27	0.38
	(1.28)	(2.53)	(15.17)	(0.31)	(0.34)	(0.34)
Job function: Kindergarten	1.03	0.86†	1.00	1.24*	0.99	1.20
	(0.08)	(0.08)	(0.12)	(0.15)	(0.13)	(0.19)
Job function: Secondary	0.94	0.86†	1.06	1.07	0.90	0.96
	(0.07)	(0.07)	(0.11)	(0.11)	(0.09)	(0.11)
Job function: Special education	0.99	1.11	1.05	0.94	0.96	0.87
	(0.08)	(0.11)	(0.13)	(0.11)	(0.12)	(0.13)
Job function: Gifted education	1.02	1.63**	1.79**	1.17	1.86**	1.70*
	(0.15)	(0.30)	(0.40)	(0.25)	(0.42)	(0.39)
2 years of experience	1.25**	1.22**	1.28**	1.17**	1.16*	1.14†
	(0.07)	(0.08)	(0.11)	(0.07)	(0.07)	(0.08)
3 years of experience	1.42**	1.39**	1.36**	1.20†	1.28*	1.40**
	(0.11)	(0.12)	(0.15)	(0.13)	(0.14)	(0.17)
4 years of experience	1.38**	1.42**	1.46**	1.22	1.33†	1.65**
	(0.14)	(0.16)	(0.19)	(0.18)	(0.21)	(0.29)
5 years of experience	1.44**	1.53**	1.90**	1.22	1.43†	1.81**
	(0.18)	(0.22)	(0.32)	(0.24)	(0.29)	(0.41)

Variable	One-year retention			Three-year retention		
	School	District	State	School	District	State
6 years of experience	1.73**	1.85**	2.47**	0.91	1.47	1.64
	(0.26)	(0.33)	(0.53)	(0.25)	(0.44)	(0.53)
7 years of experience	1.62**	2.01**	2.26**	2.82	na	na
	(0.30)	(0.44)	(0.58)	(3.48)	na	na
8 years of experience	1.38	1.73†	2.72*	na	na	na
	(0.38)	(0.55)	(1.13)	na	na	na
9 years of experience	0.60	na	na	na	na	na
	(0.52)	na	na	na	na	na
Number of teacher-year observations	19,933	19,810	19,222	11,088	11,022	10,831
Log likelihood	-9,819	-7,511	-4,982	-7,170	-6,461	-5,128

† Significant at $p < .10$; * significant at $p < .05$; ** significant at $p < .01$.
na is not applicable.
Note: Numbers in parentheses are clustered standard errors at the school level. The model includes indicator variables for teacher experience, as well as cohort, program, district, and year fixed effects.
Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table B6 is a supplement to figure 3 in the main report. The columns are student standardized test scores in math and English language arts, and the rows are key independent variables. The treatment row reports the estimated association between Believe and Prepare and student test scores. Each model was estimated by an ordinary least squares regression, and coefficients are in standard deviation units of student test scores.

Table B6. Results from the regressions on student math and English language arts achievement, 2016/17–2018/19

Variable	Math	English language arts
	Coefficient (standard error)	Coefficient (standard error)
Treatment	-0.01	-0.04*
	(0.02)	(0.02)
Student variables		
Prior math achievement	0.63**	0.22**
	(0.00)	(0.00)
Prior English language arts achievement	0.16**	0.54**
	(0.00)	(0.00)
Female	-0.01**	0.12**
	(0.00)	(0.00)
Black	-0.11**	-0.09**
	(0.01)	(0.00)
Hispanic	-0.04**	0.01
	(0.01)	(0.01)
Eligible for the National School Lunch Program	-0.06**	-0.07
	(0.01)	(0.00)
English learner student	-0.01	-0.16**
	(0.02)	(0.02)
In special education	-0.11**	-0.18**
	(0.01)	(0.01)

Variable	Math Coefficient (standard error)	English language arts Coefficient (standard error)
Grade 4	0.06**	0.07**
	(0.02)	(0.01)
Grade 5	-0.02	-0.05**
	(0.02)	(0.01)
Grade 6	0.03*	0.02*
	(0.01)	(0.01)
Grade 7	-0.05	0.03
	(0.05)	(0.02)
Grade 8	0.00	-0.03
	(0.05)	(0.02)
Changed school during the year	-0.09	-0.11*
	(0.07)	(0.05)
Repeated grade	-0.21**	-0.28**
	(0.05)	(0.03)
Teacher variables		
Black	0.00	-0.02
	(0.02)	(0.01)
Hispanic	0.01	0.04
	(0.05)	(0.03)
Female	-0.05*	0.05**
	(0.02)	(0.01)
Bachelor's degree	0.04*	-0.03
	(0.02)	(0.04)
2 years of experience	0.06**	0.02**
	(0.01)	(0.01)
3 years of experience	0.11**	0.01
	(0.01)	(0.01)
4 years of experience	0.11**	0.06**
	(0.01)	(0.01)
5 years of experience	0.11**	0.08**
	(0.02)	(0.01)
6 years of experience	0.15**	0.07**
	(0.0)	(0.01)
7 years of experience	0.22**	0.02
	(0.04)	(0.03)
Praxis I score	0.03†	0.01
	(0.01)	(0.01)
Constant	-0.23*	-0.06
	(0.08)	(0.06)
R^2	0.68	0.64
Number of student-year observations	74,510	119,516

† Significant at $p < .10$; * significant at $p < .05$; ** significant at $p < .01$.

Note: Numbers in parentheses are clustered standard errors at the classroom level. The model includes indicator variables for teacher preparation program graduation cohort and missing data, as well as program fixed effects.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table B7 is a supplement to figure 4 in the main report. The columns are teacher Praxis II scores, and the rows are independent variables. In addition to demonstrating that whether a teacher completed a preparation program that had implemented Believe and Prepare is not statistically significantly associated with Praxis II scores, the table shows a strong correlation between Praxis I and Praxis II scores in math and reading. Because the outcomes are preservice outcomes, teacher experience is not relevant and was dropped from all specifications.

Table B7. Regression coefficient and standard error estimates for analyses of the relationship between teacher Praxis II scores and whether a teacher completed a preparation program that had implemented Believe and Prepare, 2015/16–2019/20

Variable	Average	Content	Pedagogy	English	Math	Science	Social studies	Elementary	Special education
Treatment	–0.05	–0.01	–0.10	0.22	0.12	0.07	0.01	–0.10	0.43
	(0.05)	(0.07)	(0.08)	(0.19)	(0.24)	(0.21)	(0.14)	(0.10)	(0.42)
Praxis I math score	0.12**	0.19**	0.07**	0.09†	0.29**	0.13†	0.04	0.22**	0.00
	(0.02)	(0.02)	(0.02)	(0.05)	(0.07)	(0.07)	(0.06)	(0.03)	(0.06)
Praxis I reading score	0.20**	0.19**	0.22**	0.23**	0.09	0.13*	0.19**	0.19**	0.23**
	(0.02)	(0.02)	(0.03)	(0.04)	(0.05)	(0.06)	(0.06)	(0.03)	(0.06)
Praxis I writing score	0.05**	0.02	0.07**	–0.04	–0.11	0.09	–0.02	0.02	–0.03
	(0.01)	(0.02)	(0.02)	(0.07)	(0.09)	(0.08)	(0.05)	(0.02)	(0.06)
Black	–0.09†	–0.08†	–0.11	–0.09	–0.33*	0.05	–0.17	–0.10	–0.22
	(0.05)	(0.05)	(0.07)	(0.10)	(0.17)	(0.11)	(0.16)	(0.06)	(0.29)
Hispanic	–0.16	–0.11	–0.17	–0.57**	–0.07	0.21	0.13	–0.15	0.03
	(0.10)	(0.10)	(0.16)	(0.18)	(0.38)	(0.32)	(0.36)	(0.16)	(1.17)
Other race/ethnicity	0.02	–0.01	–0.01	–0.11	0.53	0.17	–0.25	0.09	0.81**
	(0.09)	(0.11)	(0.16)	(0.31)	(0.86)	(0.19)	(0.39)	(0.09)	(0.09)
Female	0.08	–0.23*	0.22**	–0.17	–0.10	–1.33**	–0.18	–0.44**	0.78†
	(0.07)	(0.11)	(0.08)	(0.20)	(0.31)	(0.35)	(0.21)	(0.16)	(0.40)
Number of teachers	2,077	1,807	2,063	456	408	395	459	1,256	159
R^2	0.30	0.27	0.23	0.25	0.24	0.28	0.22	0.27	0.28

† Significant at $p < .10$; * significant at $p < .05$; ** significant at $p < .01$.

Note: Numbers in parentheses are clustered standard errors at the teacher preparation program level. The model includes cohort and program fixed effects.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table B8 is a supplement to figure 5 in the main report. The columns are job placements, and the rows are independent variables. The results are based on logistic regressions and are reported as odds ratios. The estimated odds ratios were converted into marginal effects for figure 5 in the main report.

Table B8. Regression coefficient (odds ratios) and standard error estimates for analyses of the relationship between job placement and whether a teacher completed a program that had implemented Believe and Prepare, 2015/16–2019/20

	Placed in the school where the teacher completed a residency		Placed in a shortage area		Placed in a rural school	
Variable	Ever	First assignment	Ever	First assignment	Ever	First assignment
Treatment	1.00	0.90	0.97	1.13	0.93	1.05
	(0.17)	(0.16)	(0.19)	(0.22)	(0.14)	(0.16)
Black	0.65**	0.62**	0.92	0.92	0.56**	0.52**
	(0.10)	(0.10)	(0.17)	(0.14)	(0.06)	(0.06)
Hispanic	1.12	1.24	1.54	0.87	0.24**	0.25**
	(0.42)	(0.48)	(0.74)	(0.40)	(0.11)	(0.12)
Other race/ethnicity	0.56	0.52	1.05	0.93	0.71	0.84
	(0.20)	(0.22)	(0.45)	(0.24)	(0.19)	(0.25)
Female	0.64*	0.64*	0.87	0.80	0.94	0.96
	(0.13)	(0.14)	(0.13)	(0.12)	(0.12)	(0.12)
Number of teachers	2,993	2,951	5,181	5,233	5,594	5,571
Log likelihood	-1,590	-1,504	-1,639	-1,974	-3,388	-3,089

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Numbers in parentheses are clustered standard errors at the teacher preparation program level. The model includes cohort and program fixed effects.

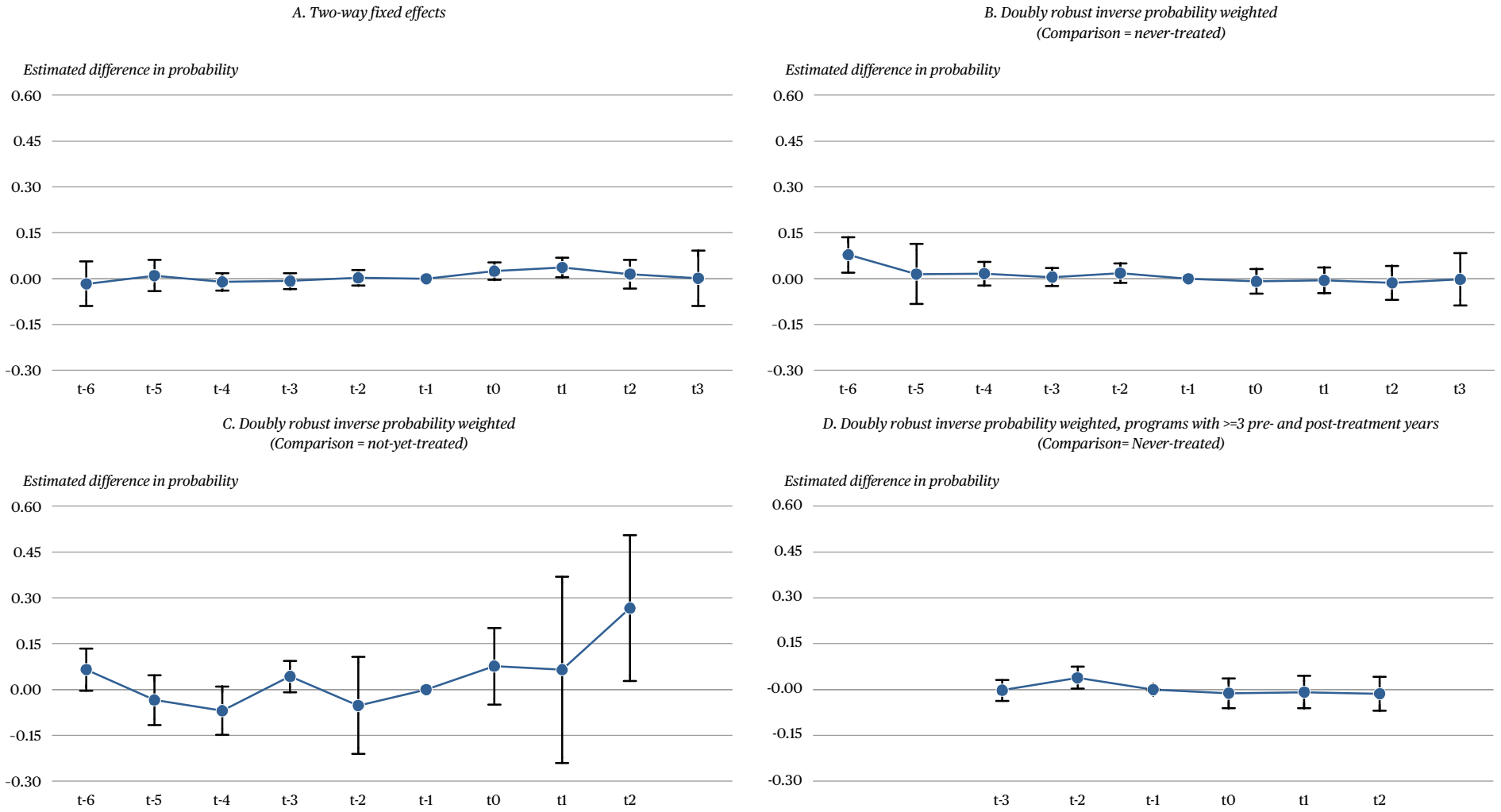
Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Appendix C. Supplemental analyses

This appendix presents additional analyses as described in the "Sensitivity analysis and robustness checks" section of appendix A.

Figure C1 depicts how the difference in the probability of staying in Louisiana public schools changed over time between teachers who completed a preparation program that later implemented Believe and Prepare and teachers who completed a program that never implemented it during the study period. The year immediately before the reform took effect, $t - 1$, is the reference period (that is, the difference in retention probability in period $t - 1$ between the two groups of teachers was set to 0). Differences in all other periods are relative to the difference in the reference period. Each panel presents a different strategy for estimating the evolution of these differences.

Figure C1. Estimated treatment effects on one-year teacher retention in Louisiana, by elapsed time since Believe and Prepare was first implemented and by estimation method and comparison group, 2012/13–2018/19



Note: The analytic samples included 17,704 teacher-year observations for panel A, 17,671 observations for panel B, 2,307 observations for panel C, and 16,086 observations for panel D. Each dot represents the estimated difference in the probability of staying in Louisiana as a teacher from year to year between teachers who completed a preparation program that had implemented Believe and Prepare and teachers who completed a program that had not implemented it. The vertical lines above and below each dot represent the 95 percent confidence intervals based on school-level clustered standard errors estimated using a multiplicative wild bootstrap procedure that takes into account the dependency across different group-time average treatment effect estimates. See Callaway and Sant'Anna (2021) for more details. The horizontal axis represents the number of years elapsed since a teacher preparation program first implemented Believe and Prepare. The last period before the first implementation, $t - 1$, is the reference period. Panel A presents estimated treatment effects using a dynamic two-way fixed effects model. The other panels present estimated effects following Callaway and Sant'Anna (2021) that use never-treated programs (panels B and D) and not-yet-treated programs (panel C) as the comparison group.
Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Tables C1 and C2 report the results of the sensitivity analyses and robustness checks for teacher retention and student English language arts achievement described in appendix A. Column 1 in each table reproduces the estimate from the main report. In table C1 the estimate is presented as a marginal effect, and column 2 uses the same specification as in column 1 but reports the treatment effect estimated using a linear probability model. In column 3 in table C1 and column 2 in table C2, the estimation is run with programs that switched in and out of treatment dropped from the sample. Because dynamic two-way fixed effects (TWFE) models and their variants are estimated as linear probability models for teacher retention and because those models can accommodate only staggered implementation (that is, treatment status is irreversible once a program implemented Believe and Prepare), the first three columns in table C1 were designed to demonstrate that these restrictions had a minimal impact on the effect presented in the main report.

The results from various TWFE models are reported in columns 4–8 in table C1 and columns 3–7 in table C2. These results are designed to be robust to staggered implementation and heterogeneous treatment effects across groups and time periods. Column 5 in table C1 and column 4 in table C2 report doubly robust estimator based on stabilized inverse probability weighting and ordinary least squares (DRIPW) estimates with never-treated programs as the comparison group. Column 6 in table C1 and column 5 in table C2 report DRIPW estimates with not-yet-treated programs as the comparison group. Column 7 in table C1 and column 6 in table C2 report DRIPW estimates when the analytic sample was restricted to programs that had data in at least three pretreatment periods and at least three post-treatment periods, with never-treated programs as the comparison group. Column 8 in table C1 and column 7 in table C2 report estimated treatment effects that were aggregated by the year in which Believe and Prepare was first implemented. In both tables *t* represents the year in which Believe and Prepare was first implemented by a program, and *t* – 1 (the last period before implementation) is the omitted time period. Each row presents the estimated effect by the time elapsed since the treatment first took effect. Coefficients for pretreatment periods provide a placebo test for the parallel trends assumption.

Table C1. Comparison of estimated average treatment effects on one-year teacher retention in Louisiana, by sample selection, model, and comparison group selection, 2012/13–2018/19

Estimated average treatment effect	(1) Main	(2) Linear probability	(3) No switch-outs	(4) Dynamic two-way fixed effects	(5) DRIPW never treated	(6) DRIPW not-yet-treated	(7) DRIPW never treated trimmed	(8) DRIPW never treated
Average effect	0.02†	0.02*	0.03**	na	na	na	na	na
	(0.01)	(0.01)	(0.01)	na	na	na	na	na
By elapsed time								
t - 6	na	na	na	-0.02	0.08*	0.07†	na	na
	na	na	na	(0.04)	(0.03)	(0.04)	na	na
t - 5	na	na	na	0.01	0.02	-0.03	na	na
	na	na	na	(0.03)	(0.05)	(0.04)	na	na
t - 4	na	na	na	-0.01	0.02	-0.07†	na	na
	na	na	na	(0.02)	(0.02)	(0.04)	na	na
t - 3	na	na	na	-0.01	0.01	0.04	0.00	na
	na	na	na	(0.01)	(0.02)	(0.03)	(0.02)	na
t - 2	na	na	na	0.00	0.02	-0.05	0.04*	na
	na	na	na	(0.01)	(0.02)	(0.08)	(0.02)	na
t	na	na	na	0.02	-0.01	0.08	-0.01	na
	na	na	na	(0.02)	(0.02)	(0.06)	(0.03)	na

Estimated average treatment effect	(1) Main	(2) Linear probability	(3) No switch-outs	(4) Dynamic two-way fixed effects	(5) DRIPW never treated	(6) DRIPW not-yet-treated	(7) DRIPW never treated trimmed	(8) DRIPW never treated
t + 1	na	na	na	0.04*	-0.01	0.06	-0.01	na
	na	na	na	(0.02)	(0.02)	(0.16)	(0.03)	na
t + 2	na	na	na	0.01	-0.01	0.27*	-0.01	na
	na	na	na	(0.02)	(0.03)	(0.12)	(0.03)	na
t + 3	na	na	na	0.00	0.00	na	na	na
	na	na	na	(0.05)	(0.04)	na	na	na
By year when first implemented								
2015	na	na	na	na	na	na	na	0.00
	na	na	na	na	na	na	na	(0.03)
2016	na	na	na	na	na	na	na	-0.02
	na	na	na	na	na	na	na	(0.03)
2017	na	na	na	na	na	na	na	-0.01
	na	na	na	na	na	na	na	(0.03)
2018	na	na	na	na	na	na	na	0.22†
	na	na	na	na	na	na	na	(0.13)
Number of teacher-year observations	19,162	19,162	17,495	17,495	17,495	4,030	16,402	17,495
p-value of joint pretrends test	na	na	na	.86	.29	.56	.20	.29
R^2	na	0.10	0.07	0.07	na	na	na	na
Log likelihood	-4,877	na	na	na	na	na	na	na

† Significant at $p < .10$; * significant at $p < .05$; ** significant at $p < .01$.

na is not applicable. DRIPW is doubly robust estimator based on stabilized inverse probability weighting and ordinary least squares.

Note: All models include the same set of covariates that is used in the main model reported in table B5. Numbers in parentheses are clustered standard errors at the school level. Models 5-8 use wild bootstrapping.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Table C2. Comparison of estimated average treatment effects on English language arts achievement in Louisiana, by sample selection, model, and comparison group selection, 2016/17-2018/19

Estimated average treatment effect	(1) Main	(2) No switch-outs	(3) Dynamic two-way fixed effects	(4) DRIPW never treated	(5) DRIPW not-yet-treated	(6) DRIPW never treated trimmed	(7) DRIPW never treated
Average effect	-0.04**	-0.03	na	na	na	na	na
	(0.01)	(0.02)	na	na	na	na	na
By elapsed time							
t - 6	na	na	-0.00	0.04	na	na	na
	na	na	(0.10)	(0.02)	na	na	na
t - 5	na	na	-0.56**	-0.02	-0.05	na	na
	na	na	(0.01)	(0.03)	(0.07)	na	na
t - 4	na	na	-0.01	0.08*	-0.06	na	na
	na	na	(0.01)	(0.04)	(0.11)	na	na
t - 3	na	na	-0.03*	-0.12**	-0.08	-0.04	na
	na	na	(0.01)	(0.03)	(0.06)	(0.04)	na

Estimated average treatment effect	(1) Main	(2) No switch-outs	(3) Dynamic two-way fixed effects	(4) DRIPW never treated	(5) DRIPW not-yet-treated	(6) DRIPW never treated trimmed	(7) DRIPW never treated
t - 2	na	na	-0.01†	0.01	-0.08	0.09	na
	na	na	(0.01)	(0.04)	(0.09)	(0.03)	na
t	na	na	-0.06**	0.10	0.07	-0.01	na
	na	na	(0.02)	(0.30)	(0.63)	(0.36)	na
t + 1	na	na	-0.05*	-0.04	0.14	-0.04	na
	na	na	(0.02)	(0.18)	(0.47)	(0.18)	na
t + 2	na	na	-0.13**	0.09	0.11	0.09	na
	na	na	(0.04)	(0.42)	(0.72)	(0.42)	na
By year when first implemented							
2015	na	na	na	na	na	na	-0.10
	na	na	na	na	na	na	(0.35)
2016	na	na	na	na	na	na	-0.06
	na	na	na	na	na	na	(0.09)
Number of teacher-year observations	119,505	107,064	119,505	108,592	64,574	78,323	108,592
p-value of joint pretrends test	na	na	.00	.00	.37	.07	.00
R^2	0.64	0.64	0.63	na	na	na	na

† Significant at $p < .10$; * significant at $p < .05$; ** significant at $p < .01$.

na is not applicable. DRIPW is doubly robust estimator based on stabilized inverse probability weighting and ordinary least squares.

Note: All models include the same set of covariates that is used in the main model reported in table B6. Numbers in parentheses are clustered standard errors at the classroom level.

Source: Authors' analysis based on data provided by the Louisiana Department of Education.

Reference

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