Final Report of the i3 Evaluation of the Collaboration and Reflection to Enhance Atlanta Teacher Effectiveness (CREATE) Teacher Residency Program

A QUASI-EXPERIMENT IN GEORGIA

February 2021

Andrew P. Jaciw

Audra Wingard

Jenna Zacamy

Li Lin

Sze-Shun Lau

Empirical Education Inc.

ACKNOWLEDGEMENTS

We are grateful to the research study participants, CREATE program staff, staff at Georgia State University (GSU), and participating schools and districts for their assistance and cooperation in conducting this research. We also appreciate Megan Toby, former Senior Research Manager, and Thanh Nguyen, Senior Research Manager at Empirical Education Inc., for their contribution to this project and report. This work has been supported by the U.S. Department of Education's Investing in Innovation program, through Award Number U411C140133. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. As the independent evaluator, Empirical Education Inc. was provided with independence in reporting the results.

ABOUT EMPIRICAL EDUCATION INC.

Empirical Education Inc. is a Silicon Valley-based research company that provides tools and services to help K-12 school systems make evidence-based decisions about the effectiveness of their programs, policies, and personnel. The company brings its expertise in research, data analysis, engineering, and project management to customers that include the U.S. Department of Education, educational publishers, foundations, leading research organizations, and state and local education agencies.

Reference this report: Jaciw, A. P., Wingard, A., Zacamy, J., Lin, L., & Lau, S. (2021). *Final Report of the i3 Evaluation of the Collaboration and Reflection to Enhance Atlanta Teacher Effectiveness (CREATE) Teacher Residency Program: A Quasi-Experiment in Georgia*. (Empirical Education Rep. No. Empirical_GSU-7031-FR1-2021-O.1). Empirical Education Inc. https://www.empiricaleducation.com/create/

Table of Contents

Chapter 1. Introduction 1
OVERVIEW OF THE CREATE RESIDENCY AND COMPARISON PROGRAMS 2
KEY RESEARCH QUESTIONS 4
Chapter 2: Study Methods 6
CHAPTER OVERVIEW 6
PARTICIPANT RECRUITMENT 6
SAMPLE 6
SCHEDULE OF MAJOR MILESTONES 7
DATA SOURCES AND COLLECTION 8
Participant Surveys 8
Georgia Department of Education Data 9
Georgia State University College of Education and Human Development Data 9
Publicly Available Data on Teacher Certification and Teaching Status 10
CREATE Program Data 10
GENERAL DESIGN 10
GENERAL APPROACH TO ANALYSIS 11
Chapter 3. Implementation Results 13
RESEARCH QUESTIONS 13
FIDELITY OF IMPLEMENTATION RESULTS 13
Descriptive Findings Related to Support, Perceived Success in Teaching, Mentorship, and Participation in CREATE Professional Learning 15
Levels of Support for Teaching 16
Levels of Success in Various Aspects of Teaching 17
Access to Mentorship 21
Together Time Meetings for the CREATE Residents 23

Chapter 4: Exploratory Impacts on Teachers' Measures of Executive Functioning, Self-Efficacy, and Commitment to Teaching....26
RESEARCH QUESTIONS....26
MEASURES....26
METHODS....27
Sample....27
Impact model....27
BASELINE EQUIVALENCE....28
RESULTS....29
Scale 1: Resilience....29
Scale 2: Mindfulness....31
Scale 3: Self-efficacy in Teaching....34
Scale 4: Commitment to Teaching....37
Scale 5: Stress Management Related to Teaching....39
Chapter 5: Confirmatory Impacts on Teacher Assessment on Performance Standards...43
RESEARCH QUESTIONS....43
MEASURES....43
SAMPLES....43
IMPACT MODEL....43
BASELINE EQUIVALENCE....44
IMPACT FINDINGS....44
TAPS Performance Standard 3: Instructional Strategies....44
TAPS Performance Standard 7: Positive Learning Environment....46
Chapter 6. Confirmatory Impacts on Student Achievement....49
RESEARCH QUESTIONS....49
MEASURES....49
SAMPLES....49
IMPACT MODEL....50

BASELINE EQUIVALENCE 51
IMPACT FINDINGS (CONFIRMATORY) 51
IMPACT FINDINGS (EXPLORATORY) 52
Chapter 7. Exploratory Impacts on Early Career Teaching Trajectories and Retention 55
RESEARCH QUESTIONS 55
DEFINITIONS OF KEY TERMS 55
MEASURES 55
Early Career Three-Year Trajectory 55
Additional Variables Used in The Analysis 56
MATCHING AND RESULTING SAMPLES 57
ANALYSIS 57
Descriptives 57
Survival Analysis 58
RESULTS 60
Descriptives 60
Main Impact Findings 64
Chapter 8. Discussion 74
References 79

Chapter 1. Introduction

Empirical Education has partnered with Atlanta Neighborhood Charter School (ANCS) to conduct an external evaluation of the Collaboration and Reflection to Enhance Atlanta Teacher Effectiveness (CREATE) teacher residency program, as part of the U.S. Department of Education's Investing in Innovation (i3) Development grant funds. CREATE seeks to raise student achievement in local high-needs schools by increasing teacher effectiveness and retention of both new and veteran educators. CREATE aims to achieve this by developing critically-conscious, compassionate, and skilled educators who are committed to teaching practices that prioritize racial justice and interrupt inequities.

The five-year quasi-experimental evaluation follows two staggered cohorts of study participants for three years each. The three years of participation in the study comprise study participants' preservice teaching year and then their first two years as full-time classroom teachers. CREATE expects their residents to spend the three study years working in Atlanta Public Schools (APS). Study participants in the comparison group may be spread throughout the state of Georgia; however, most comparison subjects did their preservice teaching year, and went on to teach, in APS or neighboring districts. The first cohort's participation in the research began in the 2015–16 school year and continued through 2017–18. The second cohort's participation began in the 2016–17 school year and continued through 2018–19 (Table 1). A third cohort was also funded through i3 through the preservice teaching year (this cohort was added after the initial study design was determined). Cohorts 1 and 2 were pooled and analyzed together under the i3 grant. Findings related to Cohort 3 in their preservice teaching are included separately in Appendix A.[1] In this report, "Year 1" refers to study participants' first year in the study, which is their preservice teaching year, as well as the CREATE residents' first year in the CREATE program. "Year 2" refers to study participants' second year in the study and first year as teachers, as well as CREATE residents' second year in the CREATE program. "Year 3" refers to study participants' third year in the study and second year as teachers, as well as CREATE residents' third and final year in the CREATE program.

TABLE 1. CREATE RESEARCH STUDY TIMELINE FOR COHORTS 1, 2, AND 3

	2015–16	2016–17	2017–18	2018–19	2019–20
Cohort 1 (main analysis)	Year 1	Year 2	Year 3		
Cohort 2 (main analysis)		Year 1	Year 2	Year 3	
Cohort 3 (supplemental analysis)			Year 1	Year 2	Year 3

Note. Light blue cells indicate groups that will be analyzed under the SEED grant, rather than the i3 grant.

This evaluation compares study participants who are in the CREATE residency program to a comparison group of study participants in Georgia State University College of Education and Human Development's (GSU CEHD) traditional credentialing program to determine if there is a positive impact of CREATE on teacher and student outcomes. Results of this study will inform teacher preparation, effectiveness, and retention policies and practices across the state of Georgia. It will also contribute to the limited but growing body of literature on residency programs for preservice teachers.

[1] The third cohort of residents will continue to participate in CREATE, along with cohorts 4–8, with funding from the Supporting Effective Educator Development (SEED) grant program. Cohorts 1 and 2 were pooled and analyzed together under the i3 grant, and cohort 3 (in its Year 2 and 3) and cohort 4–8 will be analyzed together under the SEED grant.

In this chapter, we continue with an overview to the CREATE study under the i3 grant, including the description of the CREATE and comparison group programs and key research questions. Chapter 2 presents the study methodology, participant recruitment, project milestones, and data collection sources. Chapter 3 includes the analysis of the key components of fidelity of implementation (FOI), as well as descriptive findings from the experiences of both CREATE and comparison group participants Chapter 4 provides results on the impact of CREATE on teacher self-efficacy, commitment to teaching, stress management and empathy related to teaching (hereby "stress management"), resilience, and mindfulness, as measured through surveys. Chapter 5 presents the impact of CREATE on teacher effectiveness, as measured by the Teacher Assessment on Performance Standards (TAPS). Chapter 6 includes findings related to the impact of CREATE on student achievement in mathematics and English Language Arts (ELA), as measured by the Georgia Milestones Assessment System (Georgia Milestones). Chapter 7 includes findings of the impact of CREATE on teachers' early career trajectories. We discuss the significance and implication of the findings, and offer conclusions, in Chapter 8. Given CREATE's equity-centered programming and support, we explore several key moderators, including differential effects for Black educators, where the data support such analyses.

OVERVIEW OF THE CREATE RESIDENCY AND COMPARISON PROGRAMS

At the core, CREATE is a three-year teacher residency program. Participation begins in the preservice teaching year, while residents are completing their credential at GSU CEHD. During the student-teaching phase in Year 1, residents spend time in local schools with a Cooperating Teacher, completing their preservice teaching practicum. As residents move through the three-year residency model, their role within the classroom changes. In Year 2 of the program, most CREATE residents are paired with another CREATE teacher in a single classroom. In Year 3, residents become the *sole* "teacher of record" in their own classroom. In addition to the "progressive classroom roles," CREATE residents receive support from their cohort and CREATE program team each year. One of these supports is in the form of Together Time meetings. These meetings focus on Critical Friendship (CF) and allow residents an opportunity to share and collaborate, discuss their work and dilemmas of practice, build classroom management skills, and participate in Cognitively-Based Compassion Training® (CBCT). Residents also have access to mentor teachers and the CREATE program team for support. Furthermore, residents participate in the Summer Resident Academy (SRA) the summer after graduating from GSU CEHD with their teaching credential. Through the SRA, CREATE guides residents in developing social emotional competencies, pedagogical skills, content knowledge, and the confidence they will need for success in their first year as full-time teachers. The CREATE teacher residency program has evolved since the beginning of the grant by incorporating equity-centered practices to develop critically-conscious, compassionate, and skilled educators who are committed to teaching practices that prioritize racial justice and interrupt inequities. The focus is on equipping teachers with skills that include mindfulness, compassion, communication, and a willingness to engage across differences that facilitate building meaningful relationships with students and colleagues.

CREATE is designed to strengthen novice teachers' professional knowledge. The programming described above is intended to increase teacher collaboration through mentoring and involvement in collaborative learning communities, reduce the stress that often accompanies the early years of teaching, increase collegiality and teacher support, and improve novice teachers' executive functioning and instructional planning capacity. These short-term outcomes are hypothesized to be mediators of teachers' use of research-based instructional strategies that impact students' acquisition of key knowledge and skills and the development of a well-managed, safe, and orderly environment conducive to learning.

These teacher and classroom outcomes are, in turn, conjectured to lead to positive effects on student achievement and retention of teachers, particularly teachers of color (Figure 1. CREATE Teacher Residency Program Logic Model).[2]

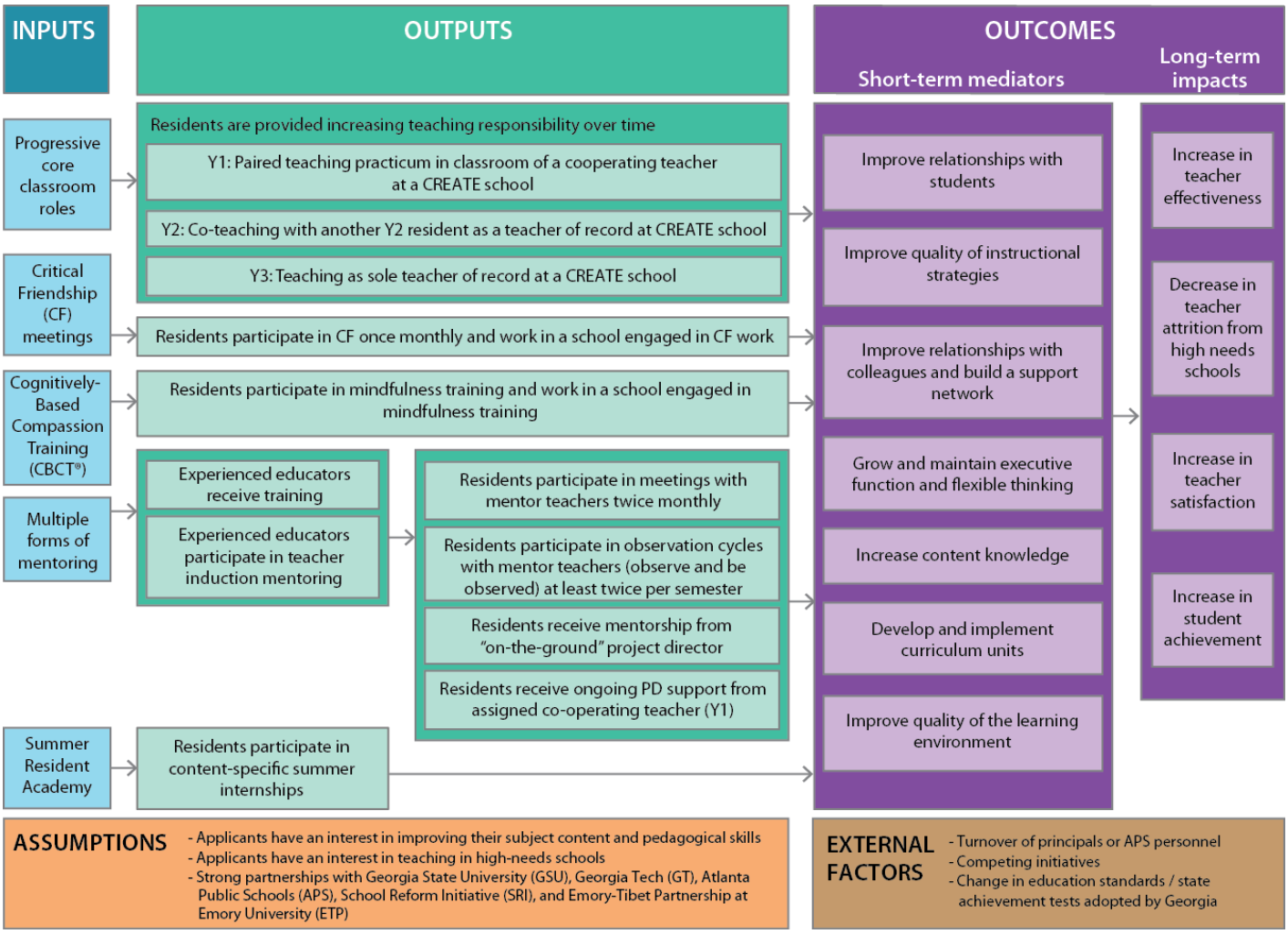

FIGURE 1. CREATE TEACHER RESIDENCY PROGRAM LOGIC MODEL

All students enrolled in GSU CEHD's teacher credentialing programs (both Early Childhood and Elementary Education and Middle and Secondary Education tracks) are invited to apply to participate in the CREATE teacher residency program. Staff members at CREATE conduct presentations at GSU CEHD, to provide students with an overview of the three-year residency program and invite them to submit an application to become a resident. These presentations usually take place in the spring and summer, setting students up to begin their residency at the beginning of their final year of study at GSU CEHD. CREATE admits students into the program based on a variety of information they provide in their application, including their interest in teaching in historically-underserved communities in Atlanta. Through these recruitment efforts, the CREATE program team is dedicated to contributing to the diversification of the teacher workforce.

[2] The CREATE program has evolved over the years. The logic model in Figure 1 reflects the CREATE program at the time of the study.

Study participants in the comparison (non-CREATE) group complete the traditional credentialing program at GSU CEHD, including GSU CEHD coursework and the in-school practicum (which CREATE residents also complete). Study participants in the comparison group may complete their practicum either in APS or another nearby district. Following graduation from GSU CEHD, comparison study participants receive no further supports from the university program. For context, according to our participant database, Cohort 1 CREATE residents commenced their practicums in one of seven CREATE schools in APS. Cohort 1 comparison group study participants completed their practicum in 63 different schools across 11 districts, including APS, over the course of the year. Cohort 2 CREATE residents commenced their practicums in one of seven CREATE schools in APS. Cohort 2 comparison group study participants completed their practicum in 59 different schools across 11 districts, including APS, over the course of the year.

KEY RESEARCH QUESTIONS

The implementation evaluation investigates the following questions.

1. Were the key components of the CREATE teacher residency logic model implemented with fidelity?
2. What is the experience of study participants in the CREATE teacher residency program and in the comparison group, specifically with regard to level of support and mentorship?

The impact evaluation of the CREATE teacher residency program addresses the following *confirmatory* research questions.[3]

3. What is the impact of CREATE on **the quality of instructional strategies** used by teachers, as measured by TAPS ratings?
4. What is the impact of CREATE on **the quality of the learning environment** created by teachers, as measured by TAPS ratings?

We measure impacts on instructional strategies and the learning environment (impact questions 3 and 4) for CREATE teachers in their first year of teaching compared to the business-as-usual teachers in their first year of teaching.

5. What is the impact of CREATE on student **mathematics** achievement in grades 4–8, as measured by the Georgia Milestones Assessment System?
6. What is the impact of CREATE on student **ELA** achievement in grades 4-8, as measured by the Georgia Milestones Assessment System?
7. What is the impact of CREATE on general **(ELA and math)** achievement of students in grades 4-8, as measured by the Georgia Milestones Assessment System?

We measure impacts on student achievement (impact questions 5, 6, and 7) for students with one year of exposure to CREATE teachers in their first year of teaching compared to students with one year of exposure to teachers in the business-as-usual group in their first year of teaching.

[3] Note, each of the five confirmatory research questions addresses a different outcome domain. No adjustments for multiple comparisons are planned. (The domains in impact questions 5 and 6 are combined in 7 but no adjustment for multiple comparisons will be necessary. This was confirmed in communication with NEi3.)

The impact evaluation of the CREATE teacher residency program also addresses the following *exploratory* research questions based on discussion with the CREATE team.

8. During their first year of teaching, what is the impact of CREATE on teacher-reported levels of self-efficacy in teaching, commitment to teaching, stress management, resilience, and mindfulness, as measured by teacher surveys?
9. Is the impact of CREATE on teacher-reported levels of self-efficacy in teaching, commitment to teaching, stress management, resilience, and mindfulness different for teachers with different baseline characteristics, including their motivation for entering teaching, confidence in general teaching skills, level of math anxiety, postsecondary GPA, and race?

We measure impacts and differential effects on survey scales (impact question 8 and 9) for CREATE teachers in their first year of teaching compared to the business-as-usual group in their first year of teaching.

10. What is the impact of CREATE on completion of the teacher preparation program at GSU CEHD and teacher retention into the first and second year of teaching for the overall sample, and for Black and non-Black educators?

Chapter 2: Study Methods

CHAPTER OVERVIEW

We conducted a quasi-experimental study to evaluate the impact of CREATE on teachers' measures of executive functioning (stress management, resilience, mindfulness), self-efficacy and commitment to teaching, teacher performance, student achievement, and teacher retention. The design compared outcomes for CREATE participants, with those of similar GSU CEHD participants who did not enroll in CREATE. This chapter provides an overview of the study methods, including participant recruitment, sample, schedule of major milestones, data sources and collection, general study design, and the general approach to analysis.

PARTICIPANT RECRUITMENT

Recruitment for Cohort 1 began in spring 2015, and recruitment for Cohort 2 began in spring 2016. Each year, we presented the research study to students who were—in their final year of GSU CEHD's teacher credentialing program—identified as eligible for participation in the research. We recruited treatment cases from the pool of students who were eligible for CREATE and who chose to join the program. We recruited comparison cases for the study from the pool of students who were eligible for CREATE but who chose to not join the program (for a variety of reasons described in the General Design section below).

In order for student teachers to be eligible for inclusion in the research study, they needed to:

- be enrolled in GSU CEHD,
- plan to teach in a public school in Georgia,
- plan to teach in an elementary or middle school, and
- expect to complete the teacher certification requirements and graduate from GSU CEHD in the spring of the first year of participation in the research.

Researchers held both in-person and virtual recruitment events for both cohorts. In the presentations, researchers provided potential study participants with information about the research study and data collection activities, and then provided them with an opportunity to ask questions. After the in-person presentations, researchers asked those interested in participating in the study to complete hard copy consent forms and return them to the researchers. A similar process occurred for the virtual presentation: a professor collected the hard copy consent forms that had been completed by interested participants and mailed them back to Empirical Education. Researchers also emailed CREATE residents who had not yet consented to the research a link to a recorded version of the recruitment presentation and an invitation to complete an online consent form.

SAMPLE

Recruitment efforts resulted in 43 CREATE residents and 99 comparison study participants who agreed to participate in the research across the two cohorts; 20 CREATE and 59 comparison study participants in Cohort 1, and 23 CREATE and 40 comparison study participants in Cohort 2. The analytic samples differ across outcomes. We describe them in detail in their respective chapters. Details about participant attrition from the study across the three years for each cohort are provided in Appendix B.

SCHEDULE OF MAJOR MILESTONES

Table 2 lists the study's major milestones for Cohorts 1 and 2.

TABLE 2. RESEARCH MILESTONES FOR COHORTS 1 AND 2

Date	Milestone
Spring 2015	Recruited Cohort 1 for CREATE residency program and research study
March 2015	Submitted application to external IRB and received exemption from full review
April–July 2015	Collected signed consent forms from and administered the baseline survey to Cohort 1
November 2015	Deployed first quarterly survey (subsequent surveys deployed in January, March, April 2016) to Cohort 1
March–August 2016	Recruited Cohort 2 for CREATE residency program and research study Collected signed consent forms from and administered the baseline survey to Cohort 2
November 2016	Deployed first quarterly survey (subsequent surveys deployed in January, March, April 2017) to both cohorts
August 2017	Finalized Memorandum of Agreement with GaDOE Submitted second interim report to CREATE program team
November 2017	Deployed first quarterly survey (subsequent surveys deployed in January, March, and April 2018) to both cohorts
October 2018	Submitted third interim report to CREATE program team
October 2018	Submitted data request to GaDOE for teacher and student outcomes, including classroom rosters, student demographic and achievement scores, and teacher TAPS ratings
November 2018	Deployed first quarterly survey to (subsequent surveys deployed in January, March, April 2019) to Cohort 2
February 2019	Submitted data request to GSU CEHD for Intern Keys ratings, Observation of Field Performance, and edTPA scores
Spring 2019–Winter 2020	Warehousing data, triangulated retention data, conducted data analysis, and drafted report
February 2021	Submitted final report

Note. IRB = Institutional Review Board; GSU CEHD = Georgia State University College of Education and Human Development; GaDOE = Georgia Department of Education; TAPS = Teacher Assessment on Performance Standards.

DATA SOURCES AND COLLECTION

This report is based on multiple sources of data that include student achievement, teacher certification, personnel records and performance, and participant surveys. We also collected participant background information through the consent process and program data from the CREATE program team in the form of rosters, attendance logs, and mentor observation logs.

Participant Surveys

Baseline Survey

After agreeing to participate in the research study, study participants in both the CREATE teacher residency program and the traditional credentialing program were invited to complete the initial baseline survey. This survey asked study participants questions about their background, motivation, perspective, and interests. Responses to this survey allowed researchers to confirm participants' eligibility for the research study, as well as informed the selection of the comparison group. Data from this survey were also used in analysis as variables for matching, as covariates in ANCOVA analysis, and as moderators in assessments of differential impacts. This survey was administered to study participants one time only, when they joined the research study.

Quarterly Surveys

Study participants in both conditions were asked to complete quarterly online surveys for the duration of the three years of the study for their respective cohort. These surveys took no more than 20 minutes each to complete, on average. Surveys included questions related to support during their student teaching year and first two years of teaching, classroom experiences, and plans for continued teaching. Appendix C includes study participant response rates for each survey.

PRIDE Teaching Environment Survey

Included in the final quarterly survey of each school year were items from the PRIDE Teaching Environment Survey. The survey assessed factors shown to be related to the likelihood that a teacher will remain in the education profession, including levels of teacher satisfaction, motivation, self-efficacy, support, career goals and intentions, school climate, and the teaching experience (Elfers et al., 2006). In the first year of participation, study participants were not yet full-time teachers and were placed at their practicum schools for varying amounts of time. Therefore, some items were adjusted to more accurately reflect the participant context.

Five Facets Mindfulness Questionnaire

Included in the final quarterly survey of each school year is the Five Facets Mindfulness Questionnaire (FFMQ). The scale is designed to measure mindfulness as represented in the psychological literature. The scale measures five facets of mindfulness: observing, describing, acting with awareness, non-judging of inner experience, and non-reactivity to inner experience. The five facets correlate with several other constructs and have incremental validity in the prediction of psychological symptoms (Baer et al., 2006).

Connor-Davidson Resilience Scale

Included in the final quarterly survey in year 2 for both cohorts was the Connor-Davidson Resilience Scale (CD-RISC).[4] The CD-RISC 10 assesses resilience and is based on the larger 25-item scale. The CD-RISC has been validated across multiple populations, countries, stressor situations, and study designs and has been used to "assess change during treatment with medication, psychotherapy, or from some other form of intervention" (Davidson & Connor, 2016).

Georgia Department of Education Data

Teacher Level Data

We collected teacher-level data from the Georgia Department of Education (GaDOE), which included TAPS ratings, gender, race, ethnicity, and termination information, if applicable. TAPS is a rubrics-based evaluation method used by GaDOE to measure Georgia public school teachers' performance on a set of designated performance standards. TAPS allows teacher effectiveness to be measured consistently throughout the state. There are ten performance standards that TAPS uses to rate teachers on a scale of 0 to 3: Level 0 is Emerging, Level I is Developing, Level II is Proficient, and Level III is Advanced. Through the programming and support it offers, CREATE aims primarily to improve teacher efficacy in two of the ten performance standards measured by TAPS, both of which are measured in this report: 1) instructional strategies (the teacher promotes student learning by using research-based instructional strategies relevant to the content area to engage students in active learning and to facilitate the students' acquisition of key knowledge and skills), and 2) positive learning environment (the teacher provides a well-managed, safe, and orderly environment that is conducive to learning and encourages respect for all) (GaDOE, 2020). The ordinal alpha, a similar measure to Cronbach's alpha, for the ten items in TAPS is 0.95, which indicates high internal consistency.

Student Level Data

Student level data collected from GaDOE include gender, age, grade level, race, ethnicity, special education status, limited English proficiency status, and Georgia Milestones ELA and mathematics scores. The Georgia Milestones assesses ELA and mathematics student achievement for students in grades 3–8, according to state-adopted content standards. The Georgia Milestones is a valid and reliable measure for student achievement in Georgia. Cronbach's alpha reliability coefficient for the Georgia Milestones ranges from 0.89 to 0.94 across all subjects, which is an adequate level of reliability for the stated goals of the assessment (GaDOE, 2019).

Georgia State University College of Education and Human Development Data

We collected teacher level data from GSU CEHD, which include study participants' practicum placements, edTPA scores (analyzed for Cohort 3 Year 1 only), and Intern Keys ratings.

edTPA is a performance-based, subject-specific, student centered, multiple measure assessment of teaching. Teacher candidates must prepare an edTPA portfolio during their student teaching practicum experience and submit it once they have completed their teaching certification program. Teacher candidates must earn a passing score on the edTPA before they can earn a teaching certificate in Georgia (GaPSC, n.d.).

---

[4] All rights reserved. Further information about the scale and terms of use can be found at www.cd-risc.com. Copyright © 2001, 2013, 2015 by Kathryn M. Connor, M.D., and Jonathan R.T. Davidson. M.D.

The teacher Intern Keys assessment (Elder et al., n.d.) is a rubrics-based evaluation that aligns directly with TAPS. University supervisors and cooperating teachers use this rubric to measure student teachers' performance on 10 state performance standards during their practicum on a rating scale of 1 to 4: Level I is Ineffective, Level II is Needs Development, Level III is Proficient, and Level IV is Exemplary. The Cronbach's alpha reliability coefficient for the teacher Intern Keys assessment is 0.90, indicating a high degree of reliability. We use the Intern Keys ratings as a baseline measure for TAPS.

Publicly Available Data on Teacher Certification and Teaching Status

Certification Data

The Educator Certification Division of the Georgia Professional Standards Commission provides a publicly available database to confirm certification status for Georgia educators (Georgia Professional Standards Commission, 2014). The database includes certification type, level, field, and issue and validity dates. This database was used to triangulate self-reported data (if needed) or fill in missing values for teacher preparation program completion.

Teaching Status

The State of Georgia provides a publicly available database to provide information on state expenditures (Open Georgia, 2008). The database includes annual salaries and travel expenses for employees of Local Boards of Education, including teachers. The research team used this information, along with a variety of other data sources, to determine teachers' teaching status.

CREATE Program Data

Researchers collected various program data from the CREATE program team in order to corroborate resident self-report survey data on FOI measures (and to report on FOI measures not addressed by resident survey data). Program data for residents include classroom placement rosters, Together Time attendance, logs for mentor meetings and observation cycles, and summer internship/academy attendance. We also collect program data for experienced educators at CREATE schools participating in CREATE activities such as attendance rosters for CF, CBCT, and mentor trainings.

GENERAL DESIGN

To address questions about the effects of CREATE on the main outcomes, as well as related questions concerning conditions for impact and differential impact, we used a comparison group design to obtain estimates of interest.[5] That is, we compared outcomes for the CREATE group with those of a matched sample of similar comparison cases. We used three design and analysis strategies to establish equivalence between CREATE and comparison groups and to reduce potential for selection bias.

The first strategy was to select a comparison group that was similar to the CREATE group. Study participants in both the CREATE and comparison groups were from a pool of students enrolled in GSU CEHD. This ensured that the comparison group participants were similar to the CREATE residents, in terms of important characteristics (including motivation to

[5] Note that the descriptive analysis of survey questions related to implementation (Chapter 3) is based on responses collected from *all* CREATE and *all* comparison teachers who completed the surveys. The sample was not limited to matched cases as described in the General Design section.

enter the teaching profession in this region of Georgia and qualifications for entering the preservice teaching program at GSU), but they chose not to join CREATE for a variety of reasons. For example, a comparison group participant may have been interested in joining CREATE but may have not wanted to teach in APS due to the distance from their home. Likewise, while they might have been committed to teaching long-term, they may not have wanted to make a three-year commitment to a specific program. Having a comparison group that was similar to the CREATE group on these factors was much more preferable than if we had selected a comparison group of study participants in colleges of education in other institutions or states.

The second strategy for ensuring a comparison group design with less potential for bias was to conduct additional matching within each cohort. This involved limiting the pool of study cases to achieve greater similarity between CREATE and comparison cases on baseline characteristics. The goal was to achieve a difference no larger than 0.25 standard deviations on any of the baseline covariates used to evaluate equivalence. This is the criterion used by the What Works Clearinghouse for assuming a tolerable level of bias that allows the study to potentially meet evidence standards with reservations (provided the covariate is also adjusted for in analysis if baseline equivalence is greater than 0.05 standard deviations). In assessing impacts on teacher surveys and retention outcomes, we assessed baseline equivalence on measures of confidence in general teaching skills, motivation to enter teaching, self-reported levels of math anxiety, and GPA at the time participants completed the baseline survey. For assessing impacts on TAPS' quality of instructional strategies and quality of teaching environment performance standards, we used baseline measures of the outcome variables. For assessing impact on student math, ELA achievement, and general achievement (math and ELA), we used student pretest scores in the corresponding subject(s) in the year before they entered classes of study teachers.

Certain adaptations of the matching methods were used with specific analyses. For example, given the small samples for analyzing impacts on TAPS ratings, we used a very basic method of trimming the sample to allow overlap between CREATE and comparison cases on the baseline measure. For analyzing confirmatory impacts on student achievement, we matched students in terms of propensity scores (i.e., estimated probabilities of being in the CREATE group) informed by both student and teacher covariates. We will note the adaptations of the methods, as necessary, and report the results of the baseline equivalence tests alongside the impact results in the respective chapters.

A third strategy for achieving greater accuracy in impact estimates was to adjust the results through analysis. Once we matched cases, we used fairly straight-forward regression-based adjustment methods. That is, our estimates of differences in outcomes between the CREATE and comparison groups adjusted for any remaining differences between the groups on baseline characteristics that can affect outcomes. The success of the methods depends more on the quality and completeness of the covariates used to make the adjustment than the sophistication of the methods (Bloom et al., 2005). We report the regression models used in the respective chapters.

GENERAL APPROACH TO ANALYSIS

After matching cases separately by cohort for each main analysis, we analyzed impacts using a series of regression-based methods. Most often, we used standard one-level linear regressions. The two exceptions were 1) the analysis of impact on student achievement, for which we used a two-level regression model (students nested in teachers) and 2) the analysis of impact on teacher retention, for which we used discrete-time survival analysis and modeled the log odds of the hazard of not being retained as the outcome. In each regression model, we included a variable indicating membership in the CREATE group or the comparison group, a variable indicating membership in Cohort 1 or Cohort 2, a series of covariates, and terms for random effects.

For each outcome, except teacher retention, we report an unadjusted and an adjusted standardized effect size—the impact estimate, divided by the pooled standard deviation of the outcome variable. To arrive at the unadjusted standardized effect size, we used a regression model that included the variable indicating cohort but excluded other baseline covariates. For the adjusted standardized effect size, we used a regression model that included the all the baseline covariates, including the variable for cohort.

For several of our analyses, we supplemented the regression methods with other approaches. For example, for the analysis of teacher TAPS ratings, we used Fisher's exact (non-parametric) test to assess the difference in ratings across conditions, given the discrete and highly non-normally distributed scores.

For all confirmatory and certain exploratory results, we tested for baseline equivalence between CREATE and comparison on specific covariates that in theory might be related to the outcome variable. This involved regressing each baseline covariate against a variable indicating membership in the CREATE or the comparison group, a variable indicating membership in Cohort 1 or Cohort 2, and terms for random effects. To determine the degree of equivalence, we examined the estimate of the regression-adjusted difference in the baseline covariate, reported in units of the pooled standard deviation of that covariate. We assessed baseline equivalence using the criteria set by the What Works Clearinghouse.

To test whether impacts were moderated by specific variables, we used the standard regression models and included a term for the interaction between the variable indicating membership in CREATE or comparison and the baseline covariate for which we were interested in examining the differential (moderated) impact. The estimate for the interaction effect indicates the added-value impact associated with each unit increase in the moderating variable. For example, it indicates the additional impact associated with being a Black educator (with group membership coded 1) relative to non-Black educators (with group membership coded 0), or the additional impact associated with each unit increase on a baseline survey measure. For outcomes based on teacher surveys (measures of executive functioning, self-efficacy, and commitment to teaching) and for teacher retention, we also examine the impacts specifically for the Black educator subgroup.

In each of the following chapters, we specify the impact model and methods in greater detail.

Chapter 3. Implementation Results

In this chapter, we present findings related to FOI across the two cohorts of study participants and provide descriptive findings from survey data about levels of support for teaching, reported success as teachers, mentorship, and participation in Together Time meetings.

RESEARCH QUESTIONS

We address the following questions concerning implementation.

- Were the key components of the CREATE teacher residency logic model implemented with fidelity?
- What is the experience of study participants in the CREATE teacher residency program and in the comparison group, specifically with regard to level of support and mentorship?

FIDELITY OF IMPLEMENTATION RESULTS

The National Evaluation of i3 (NEi3) requires that all evaluations establish key components for FOI based on the program's logic model, collect data on these components, and ultimately report on whether fidelity was met for each of the key components. We have assessed implementation fidelity for the following key components: (1) progressive core classroom roles, (2) CF work, (3) CBCT, (4) multiple forms of mentoring, and (5) paid internships. Figure 1 shows the CREATE logic model. Thresholds used for FOI can be found in the FOI matrix in Appendix D.

We assessed FOI using CREATE program rosters and resident responses on surveys. CREATE rosters were the primary data source, and we used resident self-reported attendance to fill in any cases in which there were missing data in the CREATE rosters. The FOI assessment included active CREATE residents in a given year. See Appendix B for more information about reasons residents left the CREATE program.

We present results that show which indicators within the key program components were implemented with fidelity during the three years of CREATE programming for Cohort 1 and Cohort 2 in Table 3. Cohort 3, Year 1 FOI results are reported in Appendix A.

Table 3 summarizes the FOI results for Cohorts 1 and 2 combined in each of the three years of the CREATE residency for each of CREATE's five key components, which are described in detail in Appendix E. Three of the key components of the CREATE residency program—progressive core classroom roles (Component 1); CF (Component 2); and SRA (Component 5)—were each implemented with fidelity for the years in which they were measured.

Cognitively-Based Compassion Training (Component 3) was implemented with fidelity in Year 1 and Year 3, but not in Year 2. Multiple forms of mentoring (Component 4) was not implemented with fidelity in either of the two years (years 2 and 3) in which they were measured. In Year 2, all CREATE residents had mentors who attended mentor training prior to, and during, the mentoring year. During Year 2, 94% of residents participated in at least two mentor-resident observation cycles, but only 79% of residents (instead of the targeted 95%) attended the targeted number of monthly meetings, while 18% or residents did not attend any meetings. In Year 3, all residents attended the targeted number of monthly meetings, and 94% of residents participated in at least two mentor-resident observation cycles. However, only 75% of residents (instead of 100%) had mentors who attended training prior to mentoring, and only 87% (instead of 90%) had mentors who attended training during the mentoring year.

TABLE 3. FIDELITY OF IMPLEMENTATION RESULTS FOR COHORTS 1 AND 2 COMBINED

Component	Program level threshold	Year 1	Year 2	Year 3
Progressive Core Classroom Roles	Year 1: 95% or more of residents meet fidelity on 2+ indicators Year 2: 75% or more of residents meet fidelity on 2+ indicators Year 3: 85% or more of residents meet fidelity on 2+ indicators	39/39 (100%) met fidelity on 2+ indicators **Overall: Fidelity MET**	32/34 (94%) met fidelity on 2+ indicators **Overall: Fidelity MET**	24/24 (100%) met fidelity on 2+ indicators **Overall: Fidelity MET**
Critical Friendship	Year 1: Fidelity was met for Indicator 1 and at least one other indicator Year 2: Fidelity was met for Indicator 1, and at least one other indicator Year 3: Fidelity was met for Indicator 1, Indicator 3 and at least one other indicator	Indicator 1: CREATE administrators host 2 or more institutes Indicator 2: 178/190 (94%) Indicator 3: Not measured in Y1 Indicator 4: 38/39 (97%) **Overall: Fidelity MET**	Indicator 1: CREATE administrators host 2 or more institutes Indicator 2: 161/173 (93%) Indicator 3: Not measured in Y2 Indicator 4: 31/34 (91%) **Overall: Fidelity MET**	Indicator 1: CREATE administrators host 2 or more institutes Indicator 2: 131/146 (90%) Indicator 3: 10/34 (29%) Indicator 4: 23/24 (96%) **Overall: Fidelity MET**
Cognitively-Based Compassion Training	Year 1: Fidelity was met for two indicators Year 2: Fidelity was met for two indicators Year 3: Fidelity was met for two indicators	Indicator 1: CREATE administrators host 1 or more institutes Indicator 2: 38/39 (97%) **Overall: Fidelity MET**	Indicator 1: CREATE administrators host 1 or more institutes Indicator 2: 31/34 (91%) **Overall: Fidelity WAS NOT MET**	Indicator 1: CREATE administrators host 1 or more institutes Indicator 2: 23/24 (96%) **Overall: Fidelity MET**

TABLE 3. FIDELITY OF IMPLEMENTATION RESULTS FOR COHORTS 1 AND 2 COMBINED

Component	Program level threshold	Year 1	Year 2	Year 3
Multiple forms of mentoring	Years 2 and 3: All indicators meet fidelity	Not measured in Year 1	Indicator 1: 34/34 (100%) Indicator 2: 34/34 (100%) Indicator 3: 27/34 (79%) receive a score of 2 and 6/34 (18%) receive a score of zero Indicator 4: 32/34 (94%) **Overall: Fidelity WAS NOT MET**	Indicator 1: 18/24 (75%) Indicator 2: 21/24 (87%) Indicator 3: 24/24 (100%) receive a score of 2 and none receive a score of zero Indicator 4: 22/24 (92%) **Overall: Fidelity WAS NOT MET**
Summer Resident Academy	Year 2: Indicator 1 meets fidelity	Not measured in Year 1	Indicator 1: 33/34 (97%) receive a score of at least 1 and 30/34 (88%) receive a score of 2. **Overall: Fidelity MET**	Not measured in Year 3

Note. All indicators that did meet fidelity thresholds are in green. All indicators that did NOT meet fidelity thresholds are in red.

Descriptive Findings Related to Support, Perceived Success in Teaching, Mentorship, and Participation in CREATE Professional Learning

We present descriptive findings from survey data across the three years of the study where study participants responded to questions about how supported they felt at their schools, how successful they felt in a variety of professional areas, their access to mentorship, and their level of participation in Together Time meetings.

As a developing program, CREATE has evolved since Cohort 1 began their first year. It is important to assess whether or not these programmatic changes are producing the desired results. While the summary of survey results below does not answer this question, it seeks to provide a description of how CREATE has evolved from Cohort 1 to Cohort 2. Each year of the CREATE teacher residency is somewhat unique from the other years in terms of expectations and content, so we find it most useful to look across cohorts, but within each year of the residency, to see how CREATE may or may not have evolved.

Levels of Support for Teaching

On the final quarterly survey of each year, we ask study participants to answer this question: *Overall how supported do you feel at your current practicum site* (in Year 1) or *current school* (in Years 2 and 3)?, with response options as follows: Not at all supported, Less than moderately supported, Moderately supported, More than moderately supported, Very supported. Below is a summary of the level of support reported by the CREATE and comparison groups during each year of the study.

Both CREATE residents and comparison study participants reported decreasing levels of support as they moved from Year 1 to Years 2 and 3 of the study (Figure 2 for Cohorts 1 and 2 combined). It is helpful to keep in mind that Year 1 was the study participants' student teaching year, Year 2 was study participants' first year as teachers-of-record, and Year 3 was study participants' second year as teachers-of-record, as well as CREATE residents' final year in the CREATE program. Study participants were all still students at GSU CEHD during Year 1. They spread out to their individual schools in Year 2 and took on the additional responsibilities and challenges of being a first-year teacher. They may have continued to take on even more responsibilities as second-year teachers. Though we do not know the reason for the participants' feelings of declining support levels, we think it is helpful to keep the study participants' changing and increasing responsibilities in mind. This may point to a need for CREATE to increase the level of support they offer to their Year 2 and Year 3 residents.

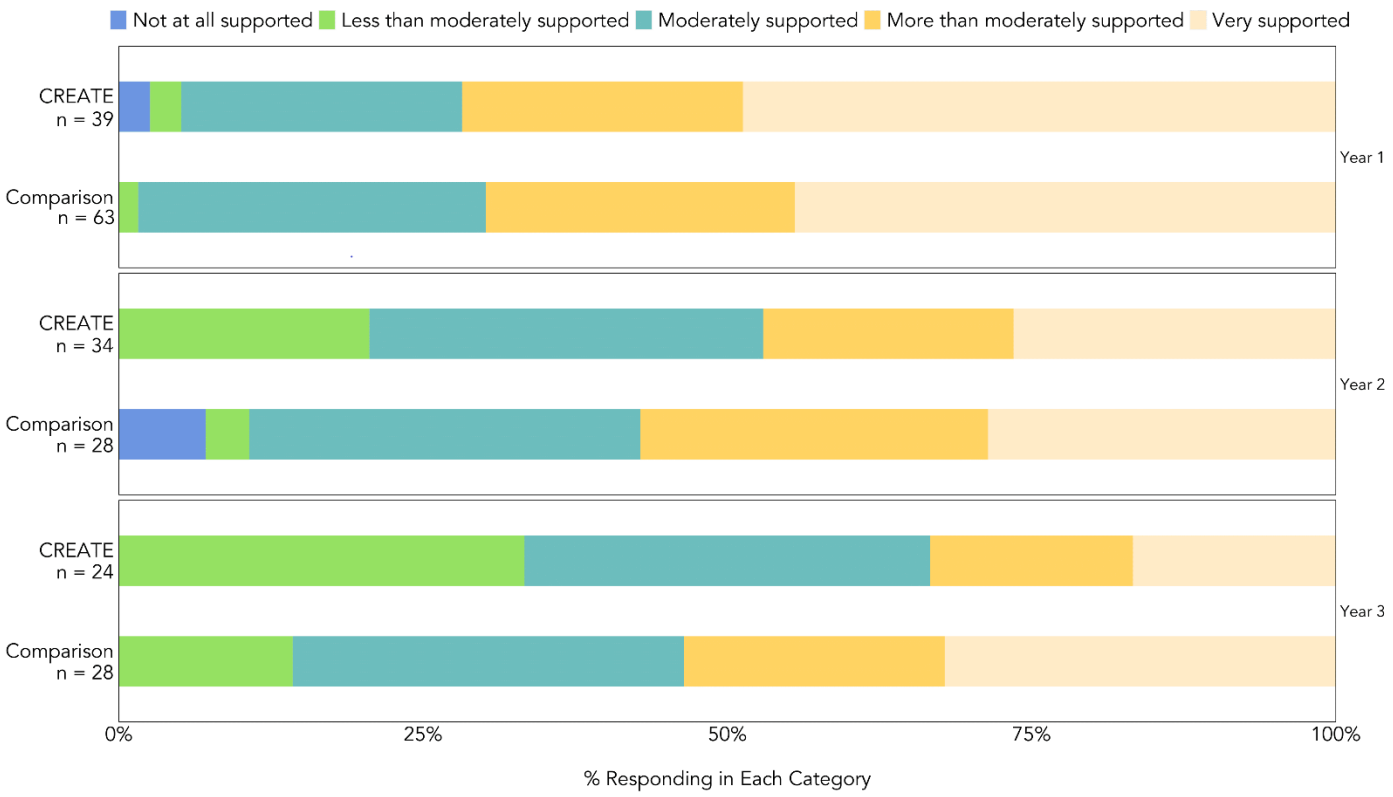

FIGURE 2. LEVEL OF SUPPORT FOR COHORT 1 AND 2 IN YEARS 1, 2, AND 3

Source: Quarterly surveys

It is also important to keep in mind that these are descriptive findings from survey responses; we did not match CREATE and comparison teachers, and did not conduct tests to examine whether the differences were statistically significant. Additionally, far more comparison teachers left the study (and left teaching) and, therefore, had no survey data reported in Years 2 and 3 (See Appendix B for more details on attrition from the study). It is also be possible that those who left teaching did not feel that they received the support they needed and those who stayed in the profession may have felt less supported (or more overwhelmed) in Years 2 and 3 and were more likely to leave the study and/or not respond to surveys.

Levels of Success in Various Aspects of Teaching

On the final quarterly survey each year, we asked study participants how successful they felt in each of the following categories (on a 5-point Likert scale from "very successful" to "not at all successful").

1. Balancing work and personal life
2. Lesson planning
3. Classroom management
4. Content knowledge
5. Pedagogical knowledge

Figure 3, Figure 4, and Figure 5, show how successful both CREATE residents and comparison study participants felt in the five categories mentioned above in Year 1, Year 2, and Year 3, respectively. Cohorts 1 and 2 are combined in these figures.

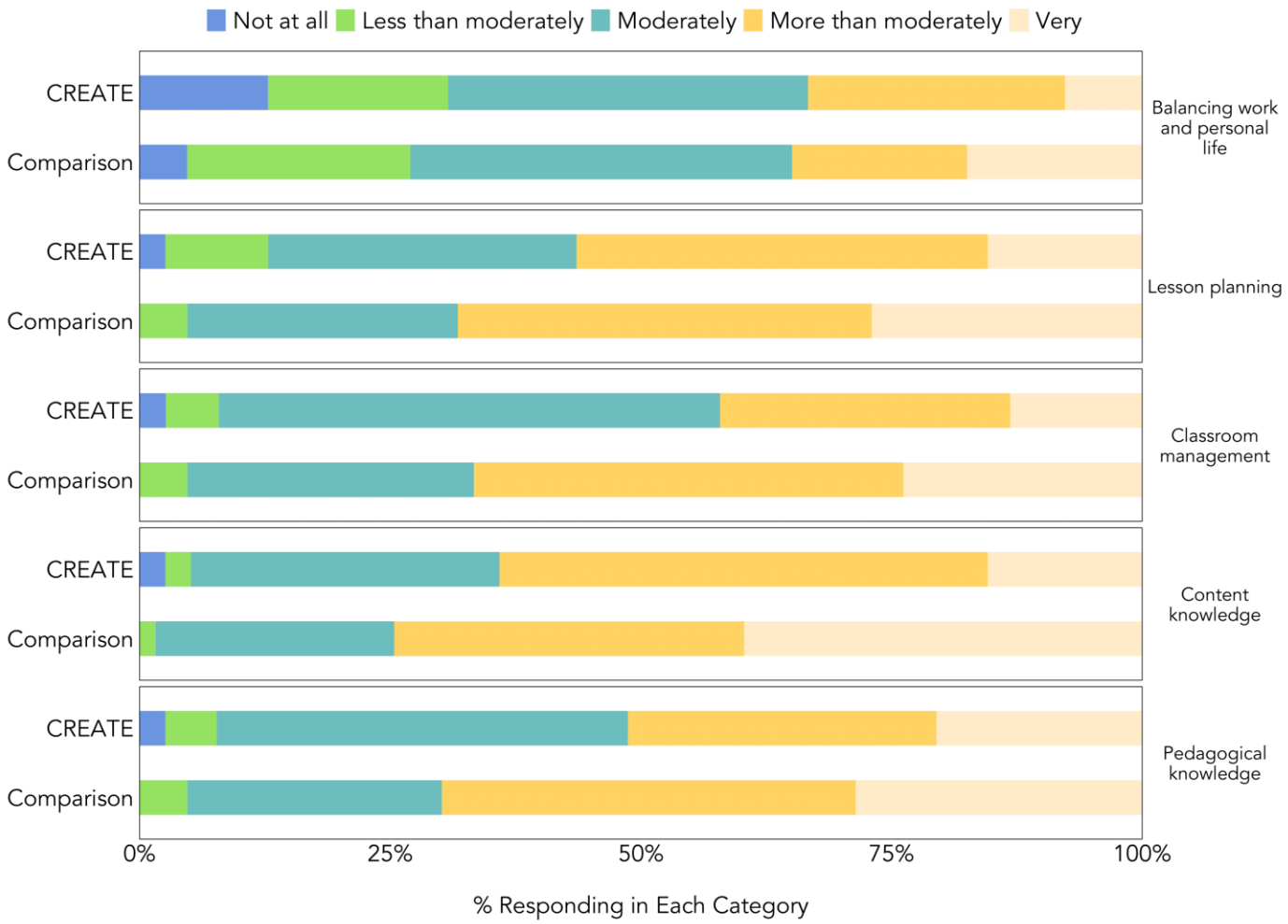

FIGURE 3. LEVELS OF SUCCESS IN VARIOUS ASPECTS OF TEACHING FOR COHORT 1 AND 2 IN YEAR 1 (STUDENT TEACHING YEAR)

Note. *N* = 39 in CREATE (except for "Classroom Management", where *N* = 38); *N* = 63 in comparison

Source: Quarterly surveys

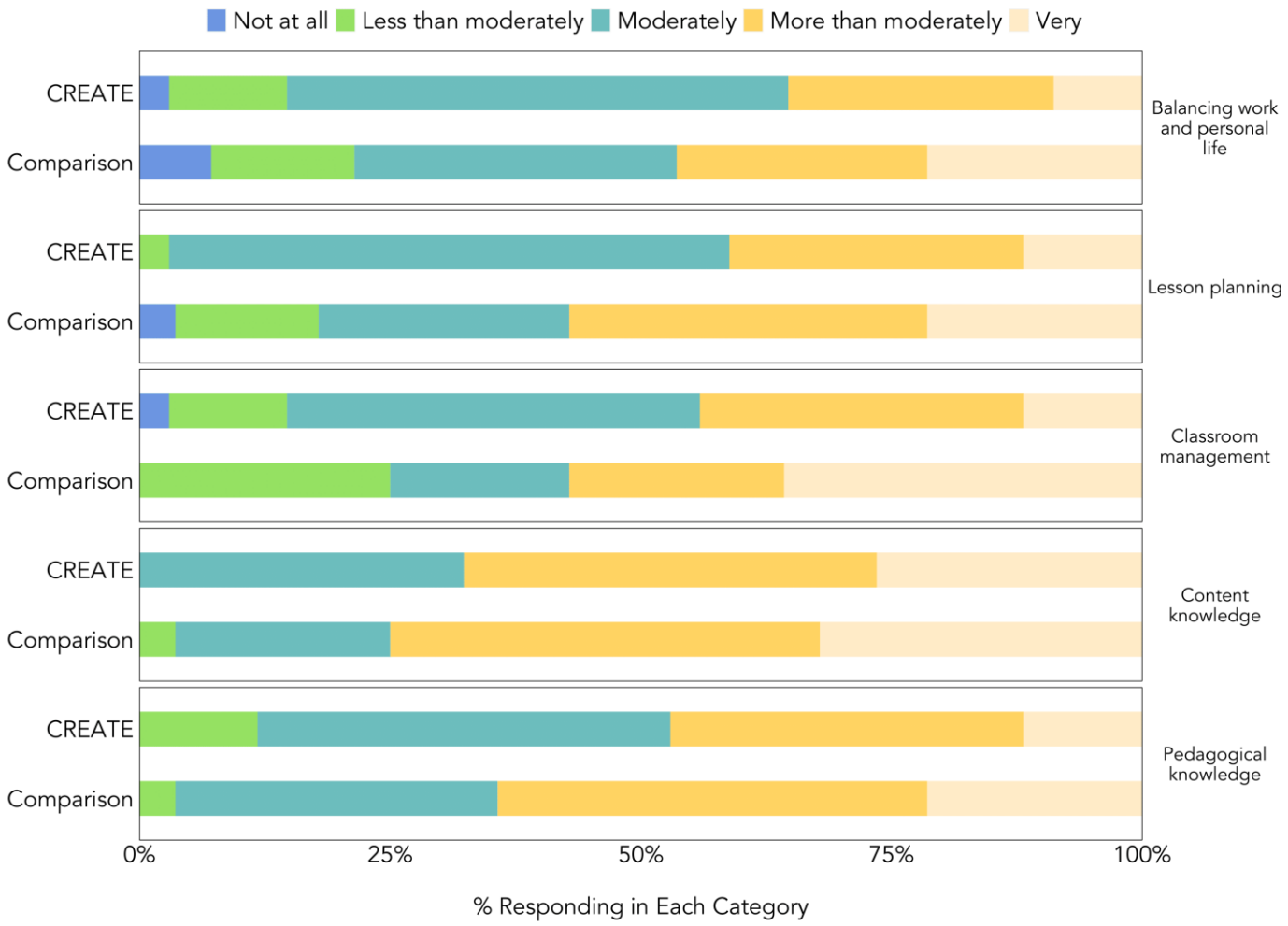

FIGURE 4. LEVELS OF SUCCESS IN VARIOUS ASPECTS OF TEACHING FOR COHORT 1 AND 2 IN YEAR 2 (FIRST YEAR AS TEACHER OF RECORD)

Note. *N* = 34 in CREATE; *N* = 36 in comparison

Source: Quarterly surveys

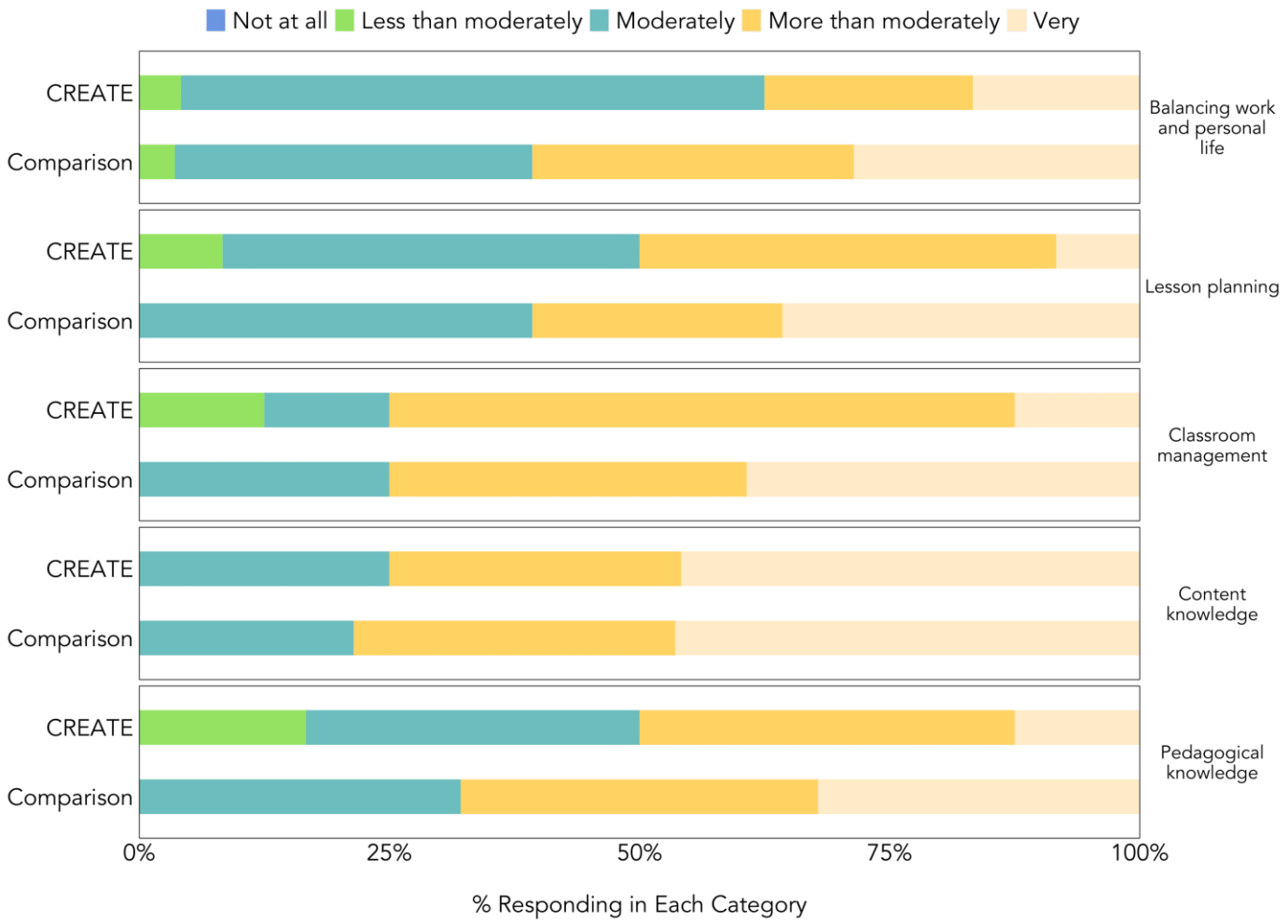

FIGURE 5. LEVELS OF SUCCESS IN VARIOUS ASPECTS OF TEACHING FOR COHORT 1 AND 2 IN YEAR 3 (SECOND YEAR AS TEACHER OF RECORD)

Note. *N* = 24 in CREATE; *N* = 28 in comparison

Source: Quarterly surveys

These descriptive findings show that both CREATE and comparison study participants feel relatively similar levels of success in the five categories, with comparison study participants, more often than not, feeling slightly more successful than the CREATE residents. However, this is a descriptive finding from survey responses; we did not test for statistical significance nor did we conduct matching. While there are a variety of reasons that comparison study participants may have reported slightly higher feelings of success, some potential (but untested) reasons may include: 1) those who are less confident in their teaching abilities (and, as a result, feel less successful) may have been more likely to sign up for the added support that CREATE provides; 2) those in CREATE have higher expectations for the level of success they should

feel given their participation in the program; 3) CREATE programming is designed to facilitate conversations among residents about what they are struggling with and how to improve. As a result, CREATE participants may have more recently and openly discussed their shortcomings before taking the survey (i.e., their responses reflect a "recency effect" of their conversations); or 4) there is higher attrition in the comparison group (see Appendix B) than in the CREATE group and many of those who left the study did so because they left teaching. It is likely that some of those that left teaching did so due to feeling unsuccessful, leaving respondents who feel more successful in the sample.

Access to Mentorship

Mentorship— a key component of the CREATE residency—is designed to support new teachers in their first two years as full-time teachers, which can be a challenging time. CREATE residents are paired with a veteran teacher at their school who provides both professional and personal support to the resident. Figure 6 and Figure 7 show that 100% of active CREATE residents in Cohorts 1 and 2 had a mentor in their first two years as full-time teachers. In contrast, 50% and 86% (in Years 2 and 3, respectively) of Cohort 1 comparison group study participants had a mentor. In Cohort 2 of the comparison group, 86% and 27% of study participants in Years 2 and 3, respectively, had a mentor. The data for these findings were collected from participant surveys and CREATE program rosters.

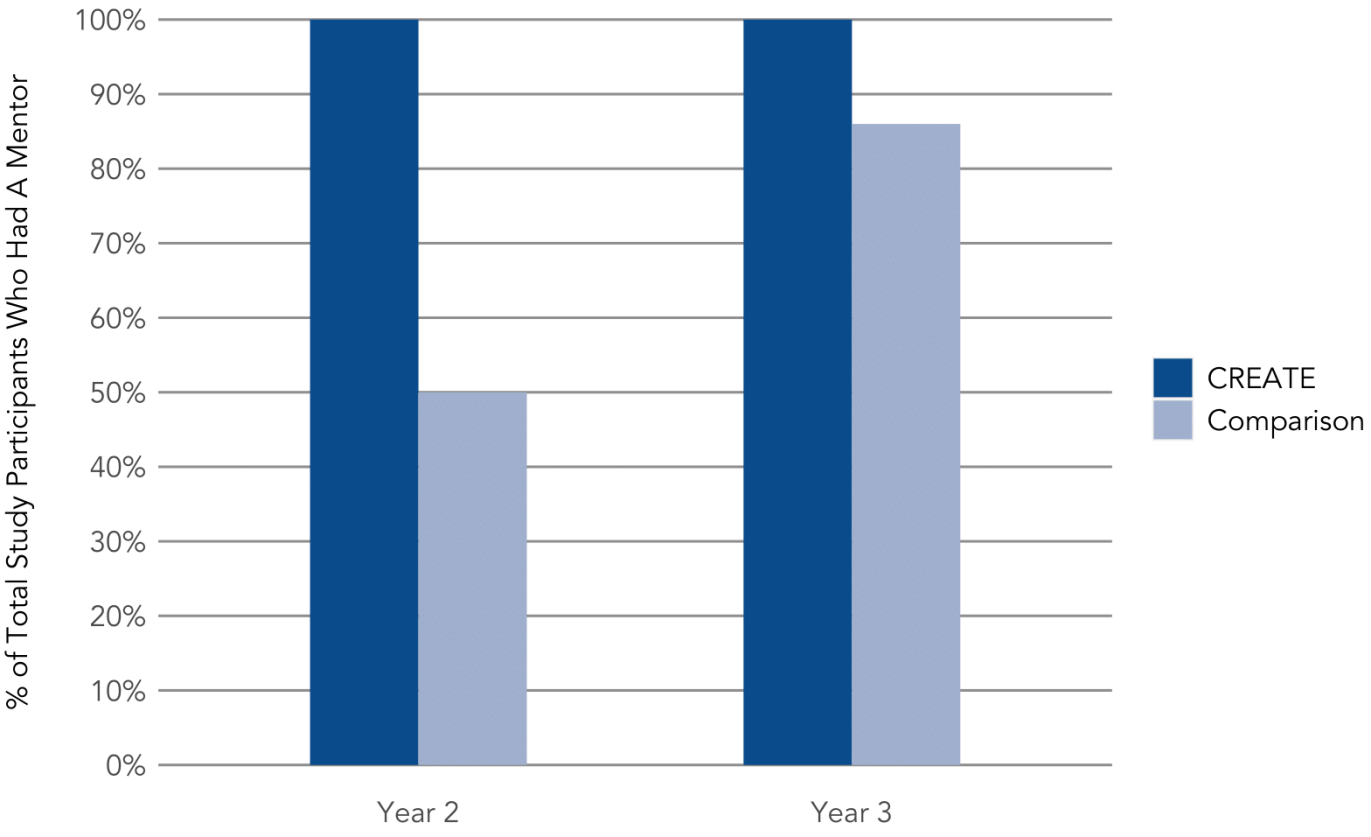

FIGURE 6. ACCESS TO MENTORSHIP FOR COHORT 1 IN YEARS 2 AND 3

Note. *N* = 10 in Y2 CREATE; *N* = 12 in Y3 CREATE; *N* = 8 in Y2 comparison; *N* = 14 in Y3 comparison

Source: Quarterly surveys and CREATE program rosters

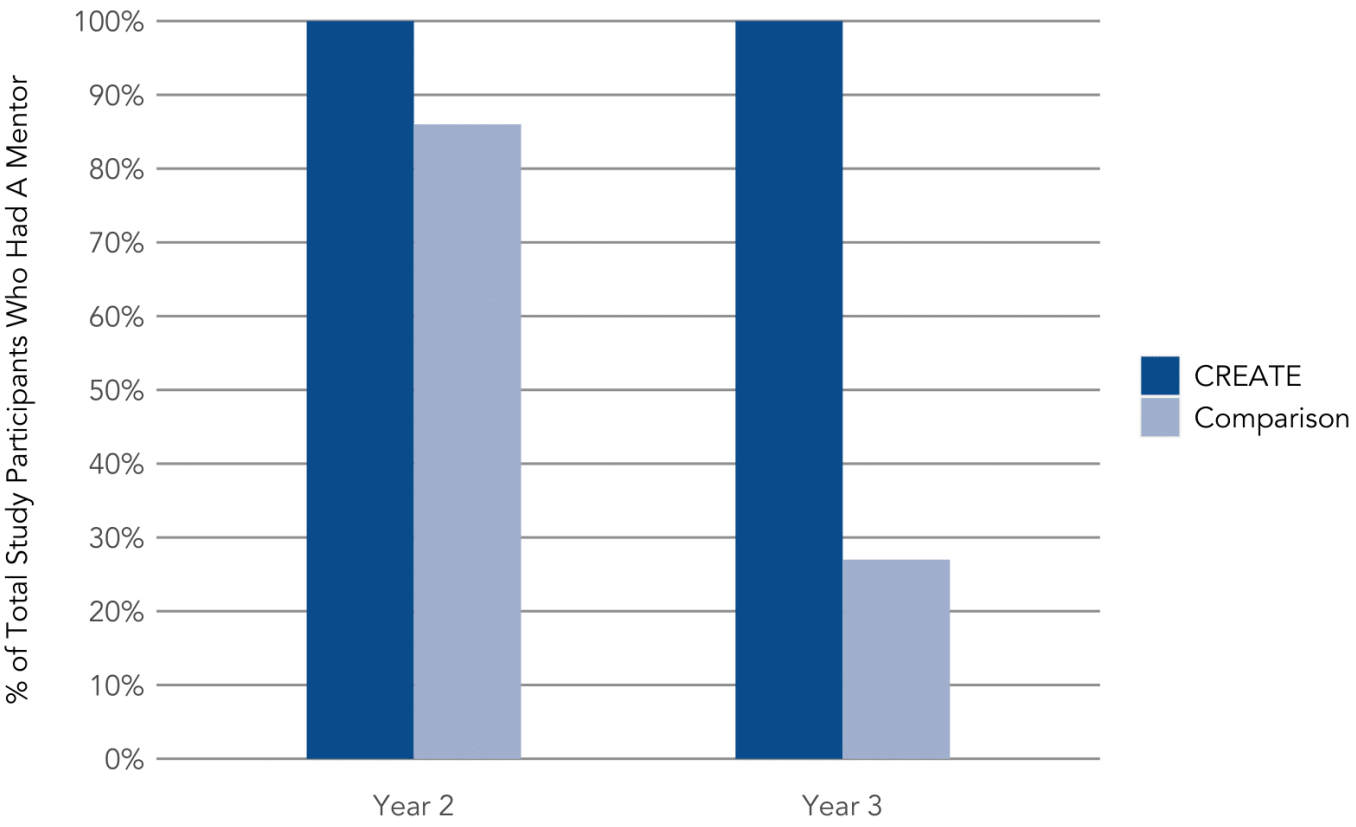

FIGURE 7. ACCESS TO MENTORSHIP FOR COHORT 2 IN YEARS 2 AND 3

Note. *N* = 15 in Y2 CREATE; *N* = 12 in Y3 CREATE; *N* = 14 in Y2 comparison; *N* = 15 in Y3 comparison

Source: Quarterly surveys and CREATE program rosters

Together Time Meetings for the CREATE Residents

Together Time meetings are another core component of CREATE's programming. As described in the introduction of this report, residents meet on a regular basis in these meetings to discuss dilemmas of practice, find support from their peers, and apply practices acquired from both Critical Friendship work and CBCT. The distribution of the number of meetings attended by CREATE residents in both cohorts during their three years in the study are shown below.

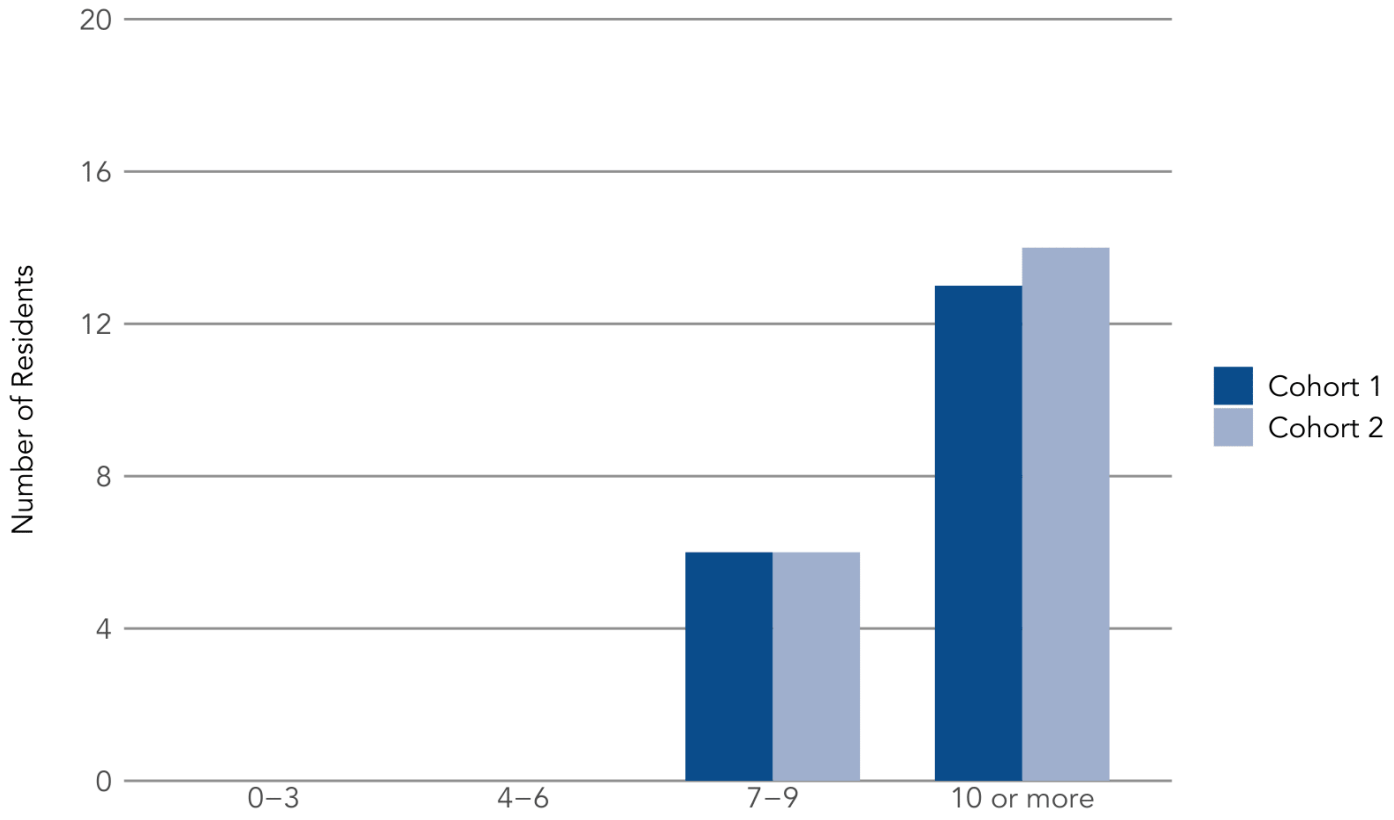

FIGURE 8. ATTENDANCE AT TOGETHER TIME MEETINGS IN YEAR 1

Note. *N* = 19 in Cohort 1; *N* = 20 in Cohort 2

Source: Quarterly surveys and CREATE program rosters

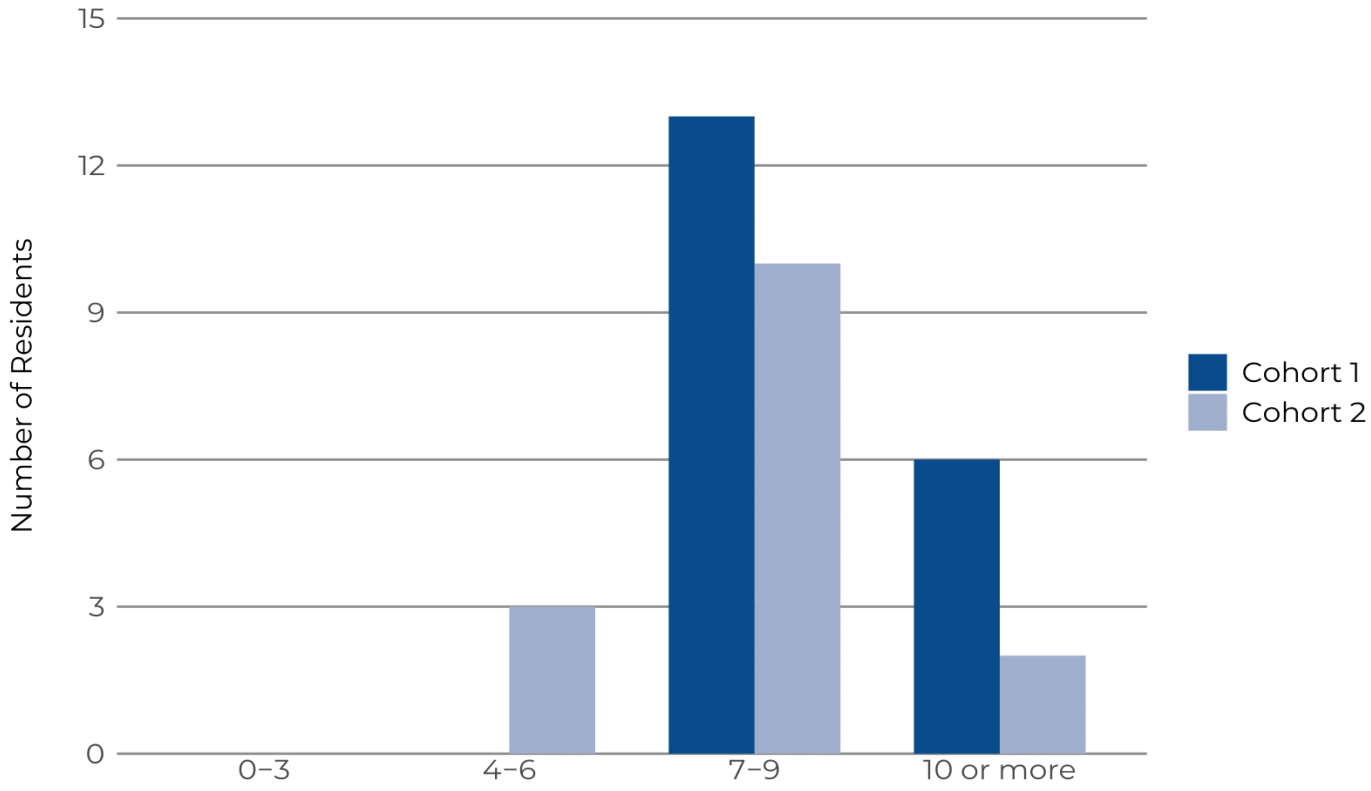

FIGURE 9. ATTENDANCE AT TOGETHER TIME MEETINGS IN YEAR 2

Note. CREATE's expectation is that residents should attend at least 7 meetings in Year 2 of CREATE residency

N = 19 in Cohort 1; *N* = 15 in Cohort 2

Source: Quarterly surveys and CREATE program rosters

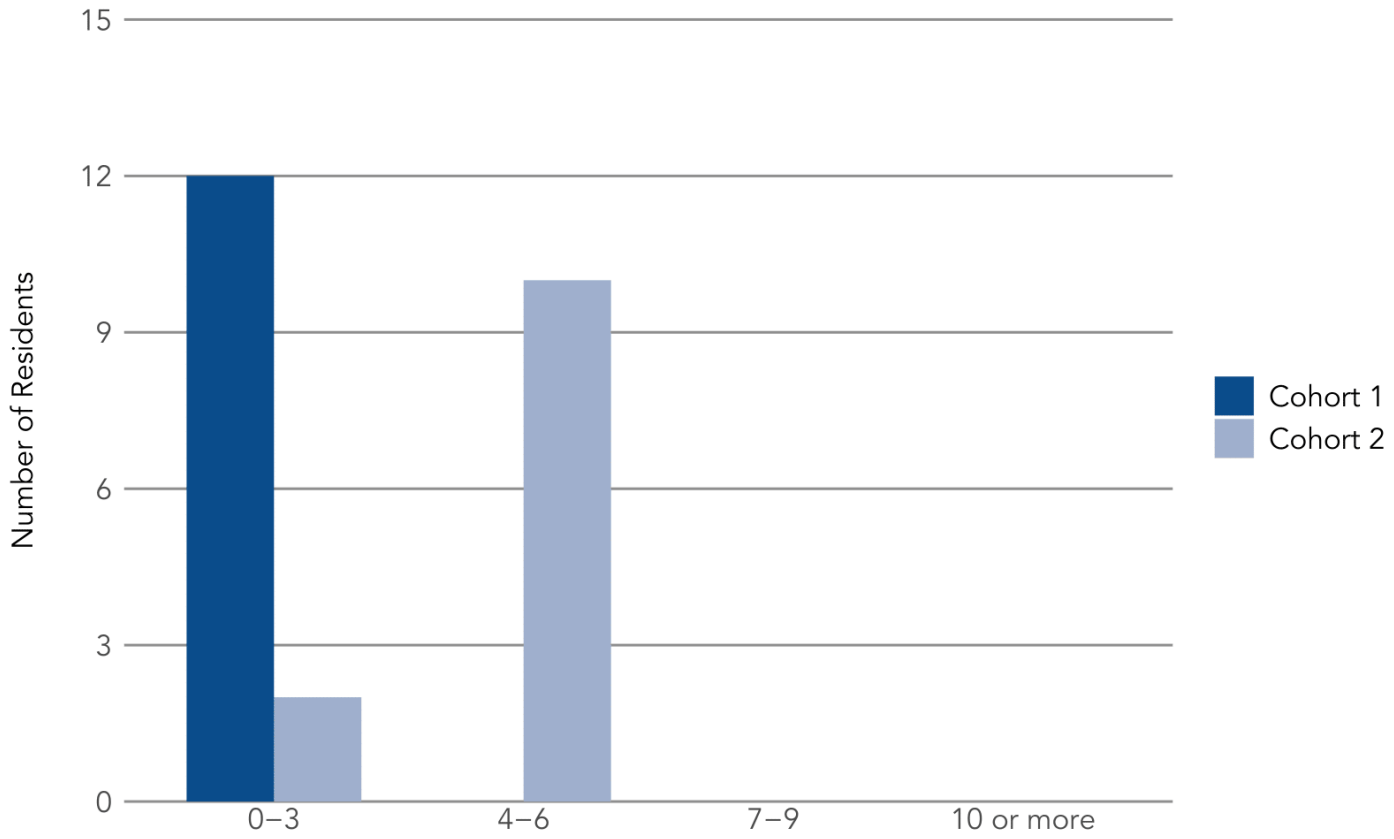

FIGURE 10. ATTENDANCE AT TOGETHER TIME MEETINGS IN YEAR 3

Note. CREATE's expectation is that residents should attend at least 3 meetings in Year 3 of the residency. There were a total of 3 meetings offered for Cohort 1 residents in Year 3 of residency and a total of 6 meetings offered for Cohort 2 residents in Year 3 of the residency.

N = 12 in Cohort 1; *N* = 12 in Cohort 2

Source: Quarterly surveys and CREATE program rosters

Chapter 4: Exploratory Impacts on Teachers' Measures of Executive Functioning, Self-Efficacy, and Commitment to Teaching

We examined average impacts on five key potential mediators. We also examined whether impacts on these potential mediators varied depending on teacher characteristics assessed at baseline. The impacts we evaluated included two cohorts of study participants.

RESEARCH QUESTIONS

We address the following questions concerning the intermediate outcomes.

- During their first year of teaching, what is the impact of CREATE on teacher-reported levels of self-efficacy in teaching, commitment to teaching, stress management, resilience, and mindfulness, as measured by the teacher surveys?
- Is the impact of CREATE on teacher-reported levels of self-efficacy in teaching, commitment to teaching, stress management, resilience, and mindfulness different for teachers with different baseline characteristics, including their incoming motivation for entering teaching, confidence in general teaching skills, level of math anxiety, current GPA, and race?

We measure impacts and differential effects on survey scales for CREATE teachers in their first year of teaching compared to the business-as-usual group in their first year of teaching.

MEASURES

The five intermediate outcomes on which we examined impacts are described below. They include important potential mediators of the effects of CREATE on more distal outcomes, such as retention of teachers in the profession. That is, if the program does not impact basic measures of executive functioning or self-regulatory behaviors, then those outcomes may not mediate longer run impacts on outcomes traditionally valued in educational policy and research. A goal of CREATE implementation is to equip teachers with skills that give them strategies to cope effectively with challenges of the profession, including potential stressors. The survey measures are meant to capture the more immediate changes. [6]

Resilience (α=.91) consisted of the 10-item CD-RISC 10 (an ordinal scale with responses ranging from 0 [not true at all] to 4 [true nearly all of the time]). Items are not specific to resilience as related to teaching (i.e., a higher score on the scale means a participant self-reports that he or she has greater resilience generally).

Mindfulness (α=.68) consisted of 12 items adopted from the Five Facets Mindfulness Questionnaire (an ordinal scale with responses ranging from 1 [never or rarely true] to 4 [very often to always true]).

Self-efficacy in teaching (α=.81) consisted of seven items from the PRIDE Teaching Environment Survey (an ordinal scale with responses ranging from 1 [not true at all] to 4 [very true]). Items addressed teachers' sense of ability to instruct, motivate students, and manage the classroom.

[6] Internal consistency reliability values are based on the study sample.

Commitment to teaching (α=.82) consisted of four adapted items from the PRIDE Teaching Environment Survey (an ordinal scale with responses ranging from 1 [not true at all] to 4 [very true]). Items addressed teachers' motivation and continued interest in teaching.

Stress management related to teaching (α=.92) consisted of six items from a researcher-developed scale (an ordinal scale with responses ranging from 1 [strongly disagree] to 5 [strongly agree]). Items addressed teachers' capacity to handle stressful situations, self-advocate, and take the perspective of students and colleagues.

The research team administered these survey scales to teachers at the end of the first year of teaching. For each scale, individual scores were obtained by averaging responses across individual items. Full descriptions of the outcome scales and the measures used as moderators in the second research question bulleted above, are provided in Appendix F.

METHODS

Sample

After limiting the sample to teachers with survey outcomes from their first year of teaching, there were 61 teachers remaining across both cohorts (28 in Cohort 1 and 33 in Cohort 2). We achieved baseline equivalence (on self-reported responses on confidence in general teaching skill, on motivation for entering teaching, and on math anxiety) for the analytic sample without the need for additional matching. Analyses are based on a sample of Cohorts 1 and 2 combined.

Impact model

The impact model used had the following form.

$$Y_i = \beta_0 + \beta_{cohort} C_i + \beta_T T_i + \sum_{p=1}^{P} X_{p,i} + \varepsilon_i \quad (1)$$

The survey score of teacher i, Y_i, was expressed as the sum of an intercept term, β_0, an effect of cohort, β_{cohort}, (C_i being coded 0 if belonging to Cohort 1, and 1 if belonging to Cohort 2), β_T, an effect of being in treatment (T_i being coded 0 if belonging to comparison, and 1 if belonging to CREATE), a series of teacher-level covariates $X_{p,i}$, and a term ε_i, representing the random deviation of a person's score from the grand mean outcome, conditional on covariates in the model.

The reported standardized effect size consists of the regression-based impact estimate divided by the pooled standard deviation of the outcome variable.

To evaluate differential impacts across the levels of moderators, we used an impact model like in Equation 1, but additionally included a term for the interaction between treatment status and the moderator. We evaluated differential effects one at a time, and with all moderator effects combined.

To determine baseline equivalence, we regressed each of three measures used to test baseline equivalence against the indicator of treatment assignment status, a dummy variable indicating cohort, and the random effect at the teacher level as in the main impact model in Equation (1). Pre-intervention measures of the outcome variables were not available; therefore, we assess baseline equivalence on three covariates that were considered to be important in influencing survey outcomes: (a) confidence in general teaching skills, (b) motivation to enter teaching, and (c) self-reported levels of math anxiety.

BASELINE EQUIVALENCE

Table 4 through Table 6 display results of tests of baseline equivalence for the analytic sample used to estimate the average impacts of CREATE on survey outcomes. Baseline equivalence with standardized mean differences of less than .25 is achieved for all three scales with cohorts combined.

TABLE 4. TESTS OF BASELINE EQUIVALENCE IN SELF-REPORTED CONFIDENCE IN GENERAL TEACHING SKILLS

	CREATE			Comparison group			Baseline difference	
Baseline measure	Sample size	Model-adjusted mean	SD	Sample size	Unadjusted mean	SD	CREATE-comparison difference	ES
Confidence in teaching (model-based approach)	33	3.933	0.548	28	4.036	0.544	-0.103	-0.189
Unadjusted sample	33	3.945	0.548	28	4.036	0.544	-0.091	-0.165

Note. Sample includes teachers from both Cohorts 1 and 2. SD = standard deviation.

TABLE 5. TESTS OF BASELINE EQUIVALENCE IN SELF-REPORTED MOTIVATION FOR ENTERING TEACHING

	CREATE			Comparison group			Baseline difference	
Baseline measure	Sample size	Model-adjusted mean	SD	Sample size	Unadjusted mean	SD	CREATE-comparison difference	ES
Motivation for entering teaching (model-based approach)	33	4.470	0.339	28	4.479	0.333	-0.009	-0.026
Unadjusted sample	33	4.479	0.339	28	4.479	0.333	0.000	0.000

Note. Sample includes teachers from both Cohorts 1 and 2. SD = standard deviation.

TABLE 6. TESTS OF BASELINE EQUIVALENCE IN SELF-REPORTED MATH ANXIETY

	CREATE			Comparison group			Baseline difference	
Baseline measure	Sample size	Model-adjusted mean	SD	Sample size	Unadjusted mean	SD	CREATE-comparison difference	ES
Self-reported math anxiety (model-based approach)	33	2.632	1.099	28	2.677	1.068	-0.045	-0.041
Unadjusted sample	33	2.615	1.099	28	2.677	1.068	-0.062	-0.057

Note. Sample includes teachers from both Cohorts 1 and 2. SD = standard deviation.

RESULTS

We report the results by outcome. For each outcome, we report the impact finding and the results of the moderator analyses.

Scale 1: Resilience

The results of the impact analysis on the resilience outcome are displayed in Table 7. We observe a covariate adjusted impact of -0.207 effect size units (p = .338). The result is not statistically significant.

TABLE 7. IMPACT OF CREATE ON TEACHER RESILIENCE (COHORT 1 AND 2 COMBINED) DURING THE FIRST YEAR OF TEACHING

	Condition	Means	Standard deviations[a]	No. of teachers	Effect size	p value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	4.011	0.736	28	-0.290	.265	-11.4%
	CREATE	3.821	0.495	33			
Adjusted effect size[b]	Comparison	4.011			-0.207	.388	-8.2%
	CREATE	3.883					

Note. CREATE defines the group receiving the CREATE program. The p values are for the corresponding impact estimates in the impact model.

[a] The unadjusted effect size is the regression-adjusted impact estimate from a model with a dummy variable indicating treatment status, a single dichotomous covariate indicating cohort, and random effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Table 8 shows the results of moderator analyses. Note that we do not show estimates of all main effects in the models, limiting them to just the treatment variable and the variable(s) for which we assess the corresponding interaction(s) with treatment.

We are interested primarily in the "additional impact" effects reported in the second half of the table. The effects are reported in the metrics of the survey scales; therefore, the "additional impact" estimates indicate the added impacts of CREATE on the resilience scale for a 1-unit increase in the moderating variable. For example, because the value of "Black educator status" is coded 0 for a non-Black educator and 1 for a Black educator, the added value impact is the additional impact of CREATE in scale score units of the outcome measure of resilience for Black educators compared to non-Black educators. Similarly, for "Moderator is Math Anxiety," the added value impact is the additional impact of CREATE in scale score units of the outcome measure of resilience associated with a 1-point increase on the Math Anxiety scale.

We observe that the impact of CREATE on resilience is greater for Black educators than non-Black educators, with an added value impact of 0.762 scale score units (p = .021). We also observe that the impact decreases by 0.663 scale score units for each unit increase in self-reported current GPA with the differential impact being marginally significant (p = .050).

To understand these differences in impact, it is useful to consider impacts for the subgroups involved. For example, for non-Black educators, the impact is -0.424 scale score units and is statistically significant (p = .039). The impact for Black educators is -0.424 + 0.762 = 0.338 scale score units (p = .175). With respect to GPA, we observe a trend of diminishing impact with a rise in GPA. For example, for individuals with a self-reported GPA of 2, the impact of CREATE is 2.136 – 2 × (0.663) = 0.810 scale score units, while for those reporting a GPA of 3, the impact of CREATE is 2.136 – 3 × (0.663) = 0.147 scale score units. CREATE may help teacher residents who have a low GPA be more resilient, but it appears the impact becomes diminished for teacher residents with a high GPA.

TABLE 8. DIFFERENTIAL IMPACTS OF CREATE TEACHER RESILIENCE (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black *N* = 59	Model 2 Moderator is Confidence in General Teaching Skill *N* = 59	Model 3 Moderator is Motivation for entering Teaching *N* = 59	Model 4 Moderator is Math Anxiety *N* = 59	Model 5 Moderator is Current GPA *N* = 59	Model 6 All Moderators included *N* = 59
Intercept	1.548 (1.111) *p* = .170	0.625 (1.326) *p* = .640	0.010 (1.630) *p* = .995	1.234 (1.173) *p* = .298	-0.364 (1.381) *p* = .793	-1.148 (2.091) *p* = .586
Main Effect of being a Black educator	-0.484 (0.263) *p* = .071					-0.269 (0.410) *p* = .514
Main Effect of confidence in general teaching skill		0.632 (0.215) *p* = .005				0.705 (0.217) *p* = .002
Main Effect of motivation for entering teaching			0.487 (0.355) *p* = .176			0.365 (0.351) *p* = .304

TABLE 8. DIFFERENTIAL IMPACTS OF CREATE TEACHER RESILIENCE (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black	Model 2 Moderator is Confidence in General Teaching Skill	Model 3 Moderator is Motivation for entering Teaching	Model 4 Moderator is Math Anxiety	Model 5 Moderator is Current GPA	Model 6 All Moderators included
Main Effect of Math Anxiety				0.056 (0.106) $p = .602$		0.019 (0.107) $p = .861$
Main Effect of Current GPA					0.450 (0.247) $p = .074$	0.237 (0.343) $p = .492$
Treatment	-0.424 (0.201) $p = .039$	1.078 (1.169) $p = .361$	2.314 (2.182) $p = .294$	0.051 (0.409) $p = .901$	2.136 (1.158) $p = .071$	3.882 (2.658) $p = .151$
Additional impact associated with being a Black educator	**0.762 (0.321) $p = .021$**					0.598 (0.454) $p = .194$
Additional impact for each unit increase in confidence in teaching skill		-0.299 (0.289) $p = .307$				-0.415 (0.299) $p = .171$
Additional impact for each unit increase in motivation for entering teaching			-0.542 (0.485) $p = .269$			-0.374 (0.494) $p = .453$
Additional impact for each unit increase in math anxiety				-0.064 (0.143) $p = .656$		-0.031 (0.142) $p = .829$
Additional impact for each unit increase in current GPA					**-0.663 (0.331) $p = .050$**	-0.247 (0.441) $p = .577$
Variance component	0.312 $p < .001$	0.340 $p < .001$	0.339 $p < .001$	0.346 $p < .001$	0.324 $p < .001$	0.323 $p < .001$

Note. Estimates are in scale score units. Standard errors are in parentheses. Moderated (differential) effects that are significant or marginally significant ($p < .10$) are bolded.

Scale 2: Mindfulness

The results of the impact analysis on levels of mindfulness are displayed in Table 9. We observe a covariate adjusted impact of 0.091 effect size units ($p = .731$). The result is not statistically significant.

TABLE 9. IMPACT OF CREATE ON LEVELS OF MINDFULNESS (COHORT 1 AND 2 COMBINED) DURING THE FIRST YEAR OF TEACHING

	Condition	Means	Standard deviations[a]	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	3.378	0.485	28	0.064	.804	2.6%
	CREATE	3.396	0.382	33			
Adjusted effect size[b]	Comparison	3.378			0.091	.731	3.6%
	CREATE	3.417					

Note. CREATE defines the group receiving the CREATE program. The *p* values are for the corresponding impact estimates in the impact model.

[a] The unadjusted effect size is the regression-adjusted impact estimate from a model with a dummy variable indicating treatment status, a single dichotomous covariate indicating cohort, and random effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

The results of the analysis of moderated impacts of CREATE on the mindfulness outcome are displayed in Table 10. We observe an increased impact of 0.462 scale units for each unit increase in self-reported confidence in teaching skills (p = .035).

To understand this differential effect, it is useful to consider impacts for the levels of the moderating variable; that is, in terms of the values of the survey scale measuring confidence in teaching skills. The estimated impacts for values of 1, 2, 3, 4 and 5 of this scale are -1.373, -0.911, -0.449, 0.013 and .475.[7] This shows a transition away from a negative impact and towards a positive impact of CREATE on mindfulness, as baseline levels of confidence in teaching increases.

TABLE 10. DIFFERENTIAL IMPACTS OF CREATE ON LEVELS OF MINDFULNESS (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black *N* = 59	Model 2 Moderator is Confidence in General Teaching Skill *N* = 59	Model 3 Moderator is Motivation for entering Teaching *N* = 59	Model 4 Moderator is Math Anxiety *N* = 59	Model 5 Moderator is Current GPA *N* = 59	Model 6 All Moderators included *N* = 59
Intercept	2.705 (0.885) *p* = .004	3.677 (0.976) *p* <.001	2.105 (1.252) *p* = .099	2.663 (0.894) *p* = .004	2.339 (1.094) *p* = .037	3.733 (1.626) *p* = .026

[7] These are estimates for the impact of CREATE on mindfulness, at each value of the confidence in teaching scale. For example, for a person scoring a 1 on the confidence in teaching scale, the impact of CREATE is -1.835 + 0.462 = -1.373. For a person scoring a 2 on the confidence in teaching scale, the impact of CREATE is -1.835 + 2*0.462 = -0.911.

TABLE 10. DIFFERENTIAL IMPACTS OF CREATE ON LEVELS OF MINDFULNESS (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black	Model 2 Moderator is Confidence in General Teaching Skill	Model 3 Moderator is Motivation for entering Teaching	Model 4 Moderator is Math Anxiety	Model 5 Moderator is Current GPA	Model 6 All Moderators included
Main Effect of being a Black educator	-0.140 (0.210) $p = .506$					-0.266 (0.318) $p = .408$
Main Effect of confidence in general teaching skill		-0.149 (0.158) $p = .352$				-0.175 (0.169) $p = .306$
Main Effect of motivation for entering teaching			0.222 (0.272) $p = .418$			0.240 (0.273) $p = .383$
Main Effect of Math Anxiety				-0.031 (0.081) $p = .705$		-0.014 (0.083) $p = .871$
Main Effect of Current GPA					0.056 (0.195) $p = .777$	-0.164 (0.266) $p = .541$
Treatment	-0.075 (0.160) $p = .643$	-1.835 (0.861) $p = .038$	1.010 (1.676) $p = .549$	-0.061 (0.312) $p = .846$	0.643 (0.917) $p = .486$	-1.083 (2.067) $p = .603$
Additional impact associated with being a Black educator	0.220 (0.256) $p = .393$					0.218 (0.353) $p = .541$
Additional impact for each unit increase in confidence in teaching skill		**0.462 (0.213)** **$p = .035$**				**0.532 (0.232)** **$p = .026$**
Additional impact for each unit increase in motivation for entering teaching			-0.222 (0.373) $p = .554$			-0.243 (0.384) $p = .530$
Additional impact for each unit increase in math anxiety				0.028 (0.109) $p = .795$		0.041 (0.110) $p = .710$
Additional impact for each unit increase in current GPA					-0.179 (0.262) $p = .497$	-0.026 (0.343) $p = .940$
Variance component	0.198 $p < .001$	0.184 $p < .001$	0.200 $p < .001$	0.201 $p < .001$	0.203 $p < .001$	0.195 $p < .001$

Note. Estimates are in scale score units. Standard errors are in parentheses. Moderated (differential) effects that are significant or marginally significant ($p < .10$) are bolded.

Scale 3: Self-efficacy in Teaching

The results of the impact analysis on self-efficacy in teaching are displayed in Table 11. We observe a covariate adjusted impact of -0.293 effect size units (p = .247). The result is not statistically significant.

TABLE 11. IMPACT OF CREATE ON SELF-EFFICACY IN TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Condition	Means	Standard deviations[a]	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	3.199	0.577	28	-0.352	.178	-13.7%
	CREATE	3.026	0.349	33			
Adjusted effect size[b]	Comparison	3.199			-0.293	.247	-11.5%
	CREATE	3.062					

Note. CREATE defines the group receiving the CREATE program. The *p* values are for the corresponding impact estimates in the impact model.

[a] The unadjusted effect size is the regression-adjusted impact estimate from a model with a dummy variable indicating treatment status, a single dichotomous covariate indicating cohort, and random effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

The results of the analysis of moderated impacts of CREATE on self-efficacy in teaching are displayed in Table 12. We observe that the impact of CREATE is greater for Black educators than non-Black educators, with an added-value impact of 0.913 scale score units (p = .016). We observe a decrease in impact of 0.327 scale units for each unit increase in self-reported math anxiety (p = .042).

To understand these differences in impact, it is useful to consider impacts for the subgroups involved. For example, for non-Black educators, the impact is -0.543 scale score units and is statistically significant (p = .021), and the impact for Black educators is -0.543 + 0.913 = 0.370 scale score units (p = .191). Next, consider impacts across levels of the math anxiety scale. The estimated impacts for values of 1, 2, 3, 4, and 5 of this scale are 0.360, 0.033, -0.294, -0.621 and -.948. This shows a transition away from a positive impact and towards a negative impact of CREATE on self-efficacy in teaching, as the baseline level of math anxiety increases.

TABLE 12. DIFFERENTIAL IMPACTS OF CREATE ON SELF-EFFICACY IN TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black *N* = 59	Model 2 Moderator is Confidence in General Teaching Skill *N* = 59	Model 3 Moderator is Motivation for entering Teaching *N* = 59	Model 4 Moderator is Math Anxiety *N* = 59	Model 5 Moderator is Current GPA *N* = 59	Model 6 All Moderators included *N* = 59
Intercept	1.197 (1.265) *p* = .348	1.521 (1.524) *p* = .323	0.643 (1.889) *p* = .735	0.564 (1.292) *p* = .664	-0.057 (1.619) *p* = .972	3.409 (2.288) *p* = .143
Main Effect of being a Black educator	-0.380 (0.299) *p* = .211					-0.910 (0.448) *p* = .048
Main Effect of confidence in general teaching skill		0.155 (0.247) *p* = .535				0.103 (0.237) *p* = .667
Main Effect of motivation for entering teaching			0.369 (0.411) *p* = .373			0.311 (0.384) *p* = .422
Main Effect of Math Anxiety				0.118 (0.117) *p* = .317		0.161 (0.117) *p* = .174
Main Effect of Current GPA					0.191 (0.289) *p* = .512	-0.611 (0.375) *p* = .110
Treatment	-0.543 (0.229) *p* = .021	-1.276 (1.344) *p* = .347	0.309 (2.528) *p* = .903	0.687 (0.450) *p* = .133	1.767 (1.357) *p* = .199	-2.121 (2.908) *p* = .470
Additional impact associated with being a Black educator	**0.913 (0.365)** ***p* = .016**					**1.313 (0.497)** ***p* = .011**
Additional impact for each unit increase in confidence in teaching skill		0.275 (0.333) *p* = .412				0.280 (0.327) *p* = .396
Additional impact for each unit increase in motivation for entering teaching			-0.108 (0.562) *p* = .849			0.008 (0.540) *p* = .988
Additional impact for each unit increase in math anxiety				**-0.327 (0.157)** ***p* = .042**		**-0.336 (0.155)** ***p* = .035**

TABLE 12. DIFFERENTIAL IMPACTS OF CREATE ON SELF-EFFICACY IN TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black	Model 2 Moderator is Confidence in General Teaching Skill	Model 3 Moderator is Motivation for entering Teaching	Model 4 Moderator is Math Anxiety	Model 5 Moderator is Current GPA	Model 6 All Moderators included
Additional impact for each unit increase in current GPA					-0.555 (0.388) $p = .158$	0.368 (0.482) $p = .449$
Variance component	0.405 $p < .001$	0.449 $p < .001$	0.455 $p < .001$	0.419 $p < .001$	0.445 $p < .001$	0.386 $p < .001$

Note. Estimates are in scale score units. Standard errors are in parentheses. Moderated (differential) effects that are significant or marginally significant ($p < .10$) are bolded.

Scale 4: Commitment to Teaching

The results of the impact analysis on levels of commitment to teaching are displayed in Table 13. We observe a covariate adjusted impact of -0.198 effect size units (p = .424). The result is not statistically significant.

TABLE 13. IMPACT OF CREATE ON COMMITMENT TO TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Condition	Means	Standard deviations[a]	No. of teachers	Effect size	p value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	3.196	0.731	28	-0.243	.338	-9.6%
	CREATE	3.000	0.673	33			
Adjusted effect size[b]	Comparison	3.196			-0.198	.424	-7.9%
	CREATE	3.058					

Note. CREATE defines the group receiving the CREATE program. The p values are for the corresponding impact estimates in the impact model.

[a] The unadjusted effect size is the regression-adjusted impact estimate from a model with a dummy variable indicating treatment status, a single dichotomous covariate indicating cohort, and random effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

The results of the analysis of moderated impacts of CREATE on commitment to teaching are displayed in Table 14. We do not observe any differential impacts for this outcome.

TABLE 14. DIFFERENTIAL IMPACTS OF CREATE ON COMMITMENT TO TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black *N* = 59	Model 2 Moderator is Confidence in General Teaching Skill *N* = 59	Model 3 Moderator is Motivation for entering Teaching *N* = 59	Model 4 Moderator is Math Anxiety *N* = 59	Model 5 Moderator is Current GPA *N* = 59	Model 6 All Moderators included *N* = 59
Intercept	1.344 (0.899) *p* = .141	0.562 (1.025) *p* = .586	0.915 (1.283) *p* = .479	1.135 (0.907) *p* = .217	0.413 (1.097) *p* = .708	-0.083 (1.685) *p* = .961
Main Effect of being a Black educator	-0.091 (0.213) *p* = .670					0.014 (0.330) *p* = .966
Main Effect of confidence in general teaching skill		0.420 (0.166) *p* = .015				0.457 (0.175) *p* = .012
Main Effect of motivation for entering teaching			0.291 (0.279) *p* = .302			0.200 (0.283) *p* = .482

TABLE 14. DIFFERENTIAL IMPACTS OF CREATE ON COMMITMENT TO TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black	Model 2 Moderator is Confidence in General Teaching Skill	Model 3 Moderator is Motivation for entering Teaching	Model 4 Moderator is Math Anxiety	Model 5 Moderator is Current GPA	Model 6 All Moderators included
Main Effect of Math Anxiety				0.055 (0.082) $p = .504$		0.035 (0.086) $p = .683$
Main Effect of Current GPA					0.190 (0.196) $p = .337$	0.148 (0.276) $p = .594$
Treatment	-0.292 (0.162) $p = .078$	1.037 (0.904) $p = .257$	0.448 (1.717) $p = .795$	0.106 (0.316) $p = .738$	1.278 (0.919) $p = .171$	2.148 (2.142) $p = .321$
Additional impact associated with being a Black educator	0.318 (0.259) $p = .225$					0.201 (0.366) $p = .585$
Additional impact for each unit increase in confidence in teaching skill		-0.300 (0.224) $p = .186$				-0.350 (0.241) $p = .152$
Additional impact for each unit increase in motivation for entering teaching			-0.136 (0.382) $p = .722$			0.035 (0.398) $p = .930$
Additional impact for each unit increase in math anxiety				-0.102 (0.110) $p = .358$		-0.081 (0.114) $p = .479$
Additional impact for each unit increase in current GPA					-0.416 (0.263) $p = .119$	-0.272 (0.355) $p = .448$
Variance component	0.204 $p < .001$	0.203 $p < .001$	0.210 $p < .001$	0.207 $p < .001$	0.204 $p < .001$	0.210 $p < .001$

Note. Estimates are in scale score units. Standard errors are in parentheses. Moderated (differential) effects that are significant or marginally significant ($p < .1$) are bolded.

Scale 5: Stress Management Related to Teaching

The results of the impact analysis on levels of stress management related to teaching are displayed in Table 15. We observe a covariate adjusted impact of 0.311 effect size units (p = .231). The result is not statistically significant.

TABLE 15. IMPACT OF CREATE ON STRESS MANAGEMENT RELATED TO TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Condition	Means	Standard deviations[a]	No. of teachers	Effect size	p value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	4.006	0.874	28	0.279	.275	11.0%
	CREATE	4.192	0.659	33			
Adjusted effect size[b]	Comparison	4.006			0.311	.231	12.2%
	CREATE	4.244					

Note. CREATE defines the group receiving the CREATE program. The p values are for the corresponding impact estimates in the impact model.

[a] The unadjusted effect size is the regression-adjusted impact estimate from a model with a dummy variable indicating treatment status, a single dichotomous covariate indicating cohort, and random effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

The results of the analysis of moderated impacts of CREATE on stress management related to teaching are displayed in Table 16. We observe that the impact of CREATE is greater for Black educators than non-Black educators, with a marginally significant added-value impact of 0.751 scale score units (p = .093).

To understand these differences in impact, it is useful to consider impacts for the subgroups involved. For example, for non-Black educators, the impact is -0.051 scale score units and is not statistically significant (p = .855), while the impact for Black educators is -0.051 + 0.751 = 0.700 scale score units (p = .042).

TABLE 16. DIFFERENTIAL IMPACTS OF CREATE ON STRESS MANAGEMENT RELATED TO TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black *N* = 59	Model 2 Moderator is Confidence in General Teaching Skill *N* = 59	Model 3 Moderator is Motivation for entering Teaching *N* = 59	Model 4 Moderator is Math Anxiety *N* = 59	Model 5 Moderator is Current GPA *N* = 59	Model 6 All Moderators included *N* = 59
Intercept	3.212 (1.521) *p* = .040	2.768 (1.788) *p* = .128	2.236 (2.199) *p* = .314	2.695 (1.539) *p* = .086	1.998 (1.906) *p* = .300	3.253 (2.886) *p* = 0.266
Main Effect of being a Black educator	-0.553 (0.360) *p* = .130					-0.819 (0.565) *p* = 0.154

TABLE 16. DIFFERENTIAL IMPACTS OF CREATE ON STRESS MANAGEMENT RELATED TO TEACHING (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING

	Model 1 Moderator is Being Black	Model 2 Moderator is Confidence in General Teaching Skill	Model 3 Moderator is Motivation for entering Teaching	Model 4 Moderator is Math Anxiety	Model 5 Moderator is Current GPA	Model 6 All Moderators included
Main Effect of confidence in general teaching skill		0.286 (0.290) p = .329				0.270 (0.300) p = 0.372
Main Effect of motivation for entering teaching			0.222 (0.478) p = .644			0.147 (0.485) p = .762
Main Effect of Math Anxiety				0.161 (0.140) p = .255		0.173 (0.147) p = .244
Main Effect of Current GPA					0.222 (0.340) p = .517	-0.319 (0.473) p = .503
Treatment	-0.051 (0.275) p = .855	0.605 (1.577) p = .703	1.629 (2.942) p = .582	0.950 (0.536) p = .083	1.940 (1.598) p = .230	1.117 (3.669) p = .762
Additional impact associated with being a Black educator	**0.751 (0.439) p = .093**					1.007 (0.627) p = .115
Additional impact for each unit increase in confidence in teaching skill		-0.088 (0.390) p = .822				-0.134 (0.412) p = .746
Additional impact for each unit increase in motivation for entering teaching			-0.307 (0.655) p = .641			-0.187 (0.682) p = .785
Additional impact for each unit increase in math anxiety				-0.265 (0.187) p = .163		-0.270 (0.195) p = .174
Additional impact for each unit increase in current GPA					-0.487 (0.456) p = .291	0.252 (0.608) p = .680
Variance component	0.585 p <.001	0.618 p <.001	0.616 p <.001	0.595 p <.001	0.617 p <.001	0.615 p <.001

Note. Estimates are in scale score units. Standard errors are in parentheses. Moderated (differential) effects that are significant or marginally significant (p < .10) are bolded.

The results of this section are exploratory. We have not applied multiple comparison adjustments to the findings; we expect some effects to reach statistical significance by chance alone. However, the findings concerning moderated impact for Black educators are noteworthy. Specifically, we observed significant or marginally significant positive differential impacts on teacher resilience, self-efficacy in teaching, and stress management related to teaching.

Table 17 shows the adjusted means of outcomes across the five scales for Black and non-Black educators in the CREATE and comparison groups. Figure 11 is a graph of the same values. A trend we observe across the scales is that among Black educators, CREATE members consistently score higher than the comparison group on the five intermediate outcomes. The same trend is not observed among non-Black educators.

TABLE 17. ADJUSTED MEANS OF SURVEY OUTCOMES IN CREATE AND COMPARISON GROUPS FOR BLACK AND NON-BLACK EDUCATORS

Adjusted means	Resilience	Mindfulness	Self-Efficacy	Commitment	Stress management
Comparison – non-Black	1.548	2.705	1.197	1.344	3.212
Comparison – Black	1.064	2.565	0.817	1.253	2.659
Treatment – non-Black	1.124	2.63	0.654	1.052	3.161
Treatment - Black	1.402	2.71	1.187	1.279	3.359

Resilience ranges from 0 [not true at all] to 4 [true nearly all of the time]; a higher score indicates greater resilience.

Mindfulness ranges from 1 [never or rarely true] to 4 [very often to always true]); a higher score indicates greater mindfulness.

Self-efficacy in teaching ranges from 1 [not true at all] to 4 [very true]); a higher score indicates greater self-efficacy in teaching.

Commitment to teaching ranges from 1 [not true at all] to 4 [very true]); a higher score indicates greater commitment to teaching.

Stress management related to teaching ranges from 1 [strongly disagree] to 5 [strongly agree]); a higher score indicates greater stress management.

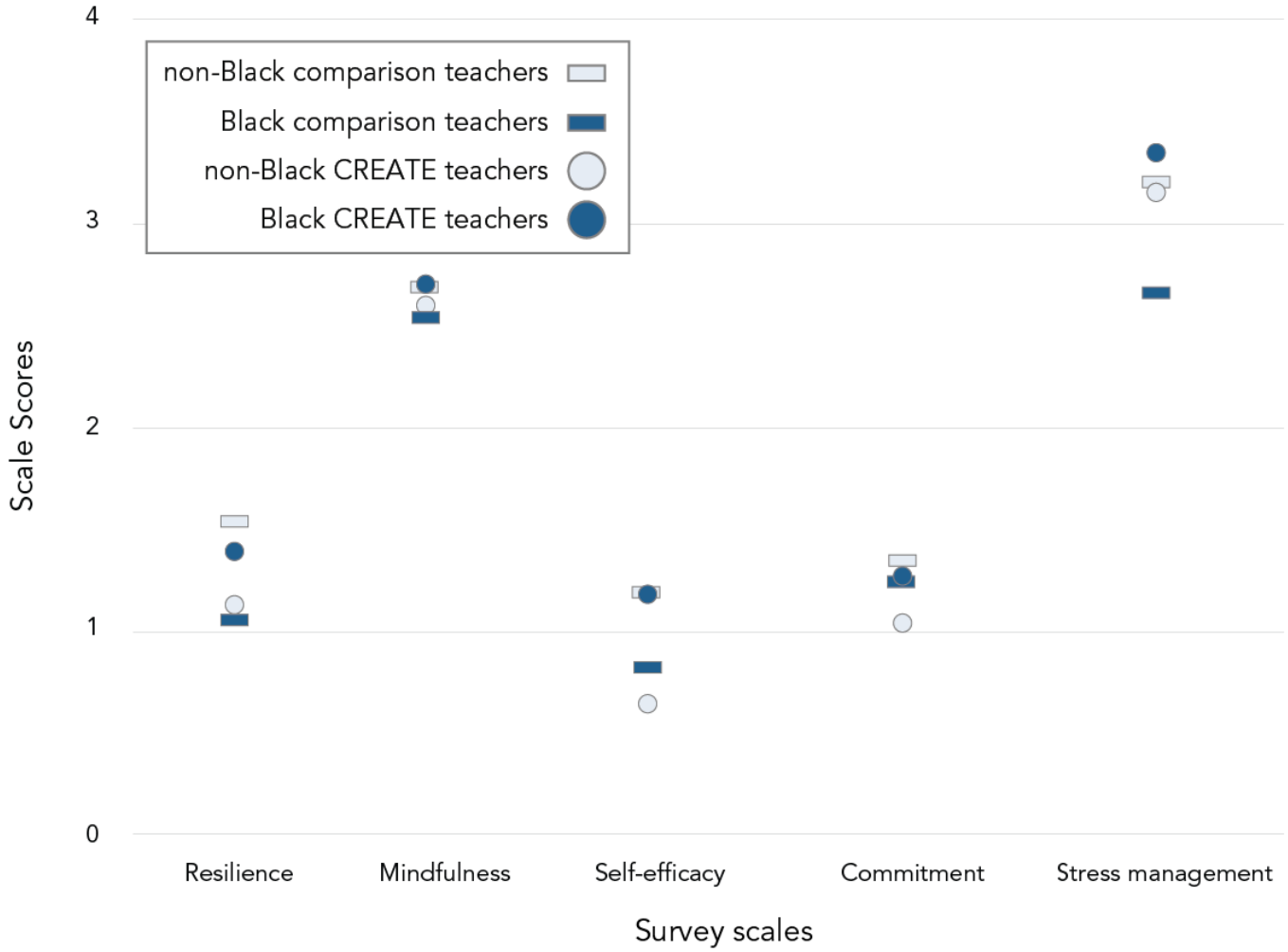

FIGURE 11. ADJUSTED MEANS OF SURVEY OUTCOMES IN CREATE AND COMPARISON FOR BLACK AND NON-BLACK EDUCATORS

Note. Impacts from outcomes scales are below. Impacts are reported in scale score units.

Resilience impacts are -0.424 (p = .039) for non-Black educators, and 0.338 (p = .175) for Black educators

Mindfulness impacts are -0.075 (p = .643) for non-Black educators, and 0.146 (p = .460) for Black educators

Self-efficacy in teaching impacts are -0.543 (p = .021) for non-Black educators, and 0.370 (p = .191) for Black educators

Commitment to Teaching impacts are -0.292 (p = .078) for non-Black educators, and 0.026 (p = .895) for Black educators

Stress Management impacts are -0.051 (p = .855) for non-Black educators, and 0.700 (p = .042) for Black educators

In this chapter, we presented results of the analysis of the impacts of CREATE on intermediate teacher outcomes assessed through surveys. While we did not observe impacts on any of the five measures for matched samples as a whole, we did find a consistent pattern of positive impacts for the subsample of Black educators. While these analyses are exploratory, they raise the possibility that downstream effects of CREATE may be mediated through different mechanisms for Black educators compared to non-Black educators. The positive impacts of CREATE on stress management related to teaching for Black educators is especially noteworthy.

Chapter 5: Confirmatory Impacts on Teacher Assessment on Performance Standards

In this chapter, we address the confirmatory impacts of CREATE on 1) the quality of instructional strategies, and 2) quality of the learning environment, both measured by the TAPS ratings. Assessing impacts on teacher effectiveness is important for determining their potential to mediate impacts on the more distal outcome of student achievement.

RESEARCH QUESTIONS

The impact evaluation of the CREATE teacher residency program addresses the following two *confirmatory* research questions regarding TAPS ratings.

- What is the impact of CREATE on **the quality of instructional strategies** used by teachers, as measured by TAPS ratings?
- What is the impact of CREATE on **the quality of the learning environment** created by teachers, as measured by TAPS ratings?

We measure impacts on instructional strategies and the learning environment for CREATE teachers in their first year of teaching compared to the business-as-usual teachers in their first year of teaching.

MEASURES

The GaDOE provided us with teacher-level data, including TAPS ratings, gender, race, ethnicity, and termination information, if applicable. GSU provided teacher-level data, including study participants' practicum placements and teacher Intern Keys ratings, which were used as the baseline measure for TAPS. More details about the data used in this analysis are available in the "Data Sources and Collection" section in chapter 2 of this report.

SAMPLES

Tables G1 and G2 in Appendix G provide details concerning the sample of teachers in the analysis of TAPS outcomes. They list the number of teachers who agreed to the study, agreed to data collection, and for whom Intern Keys ratings (baseline) and TAPS ratings (outcomes) were available for analysis. To be included in analysis, study participants had to have baseline and outcome ratings, and be matched in terms of their baseline ratings.

IMPACT MODEL

The impact model consisted of a teacher-level linear regression of the following form:

$$Y_i = \beta_0 + \beta_{cohort} C_i + \beta_T T_i + \sum_{p=1}^{P} X_{p,i} + \varepsilon_i \quad (2)$$

The rating of teacher i, Y_i, was expressed as the sum of an intercept term, β_0, an effect of cohort, β_{cohort}, (C_i being coded 0 if belonging to Cohort 1, and 1 if belonging to Cohort 2), an effect of being in treatment (T_i being coded 0 if belonging to comparison, and 1 if belonging to CREATE), a series of teacher-level covariates $X_{p,i}$ and terms for random deviations of ratings at the teacher level from the grand mean of those ratings conditional on covariates in the model ε_i.

Given the small sample sizes, we applied specific algorithms for covariate selection and described them in Appendix G under the heading "Impact Analysis."

The reported standardized effect size consists of the regression-based impact estimate in the numerator and the pooled standard deviation of the outcome variable in the denominator. We attempted to compute the effect size using the Cox index, using the cumulative log odds of responses across categories; however, in all cases, the estimation software gave the message that the maximum likelihood estimate may not or does not exist. We suspect this is due to the data being sparsely distributed across most of the response categories. We calculated the Cox index after dichotomizing the outcome (responses in the lower two levels of ordinal responses were scored 0, and those in the upper two level were scored 1) for impact on instructional strategies[8]. Given the categorical nature of the data and very few counts within certain cells in the cross-tabulation between condition and rating level, we also conducted Fisher's exact test to evaluate if we can reject the null hypothesis of no difference between conditions in the proportions of responses across rating categories.

BASELINE EQUIVALENCE

To evaluate baseline equivalence, we included a model where we regressed the Intern Keys baseline rating against an indicator of treatment status and a dummy variable to indicate cohort. The standardized effect size was -0.073 for the Intern Key Instructional Strategies ratings and -0.192 for the Positive Learning Environment ratings (baseline measures). All teachers were novices and, therefore, were perfectly matched on years of experience teaching (none). We do not have information available about the baseline achievement of students in the classes during the placement year since student teachers first became teachers of record in their second year of CREATE.

IMPACT FINDINGS

TAPS Performance Standard 3: Instructional Strategies

Table 18 shows the counts of teachers by response category for the Intern Keys and the TAPS instructional strategies performance standard by condition and by cohort. The samples were limited to cases with non-missing baseline ratingsand outcome ratings, with non-missing values for several covariates used in the impact analysis, and for whom we were able to establish baseline equivalence.

TABLE 18. COUNTS FOR EACH RATING ON THE INTERN KEYS AND TAPS INSTRUCTIONAL STRATEGIES (*N* = 27)

	CREATE	Comparison
Counts per condition and cohort		
Cohort 1	6	9
Cohort 2	8	4
Total	14	13

[8] We did not calculate the Cox index for the quality of learning environment scale because no comparison cases fell into the lower level, which resulted in an undefined value for the odds ratio associated with achieving scores at that level. The odds ratio is required to compute the Cox index.

TABLE 18. COUNTS FOR EACH RATING ON THE INTERN KEYS AND TAPS INSTRUCTIONAL STRATEGIES (*N* = 27)

	CREATE	Comparison
Counts for each rating (cohorts 1 and 2 combined) on the Intern Keys (baseline measure)		
Level 1	0	0
Level 2	7	5
Level 3	5	7
Level 4	2	1
Counts for each rating (cohorts 1 and 2 combined) on TAPS instructional strategies (outcome)		
Level 0	0	0
Level 1	2	1
Level 2	12	12
Level 3	0	0

Note.
Intern Key ratings: Level I = Ineffective; Level II = Needs Development; Level III = Proficient; Level IV = Exemplary

TAPS ratings: Level 0 = Emerging; Level I = Developing; Level II = Proficient; Level III = Advanced

Tables G1 and G2 in Appendix G give a detailed accounting of cases and how we arrived at the final analytic samples. Across both cohorts, 15 CREATE and 16 comparison teachers had both Intern Keys ratings (baseline) and TAPS ratings (outcomes), and 14 and 13 were retained in the two conditions, respectively, for analysis. The approach to matching is described in Appendix G under "Establishing Baseline Equivalence."

Ratings on the instructional strategies performance standard of TAPS were highly uniform. A possible reason for this is that raters (i.e., school principals) were reluctant to give new teachers extreme values of ratings, especially in a high-stakes environment where ratings carry consequences. This has three implications. First, the results provide no opportunity to further parse variability in outcomes through moderator analyses; that is, we would not be able to tell if impact varies based on incoming characteristics of teachers, as we are able to do with survey outcomes. Second, with very small counts and expected values for counts in several cells, logistic regression models and chi-squared tests will not yield reliable estimates. Third, the standard deviations in outcomes will be very low and, therefore, highly influential of the standardized effect size.

The results of the impact analysis are displayed in Table 19. There was no statistically significant effect of CREATE on the instructional strategies professional standard of TAPS (p = .221). As noted above, the magnitude of the effect size reflects the low standard deviation associated with near-uniform ratings of teacher instructional strategies on the TAPS performance standards. Fisher's exact test yielded results that were consistent with these. We do not reject the null hypothesis of no difference between conditions in the proportions of responses across ratings categories (p = .404).

TABLE 19. IMPACT OF CREATE ON INSTRUCTIONAL STRATEGIES (COHORT 1 AND 2 COMBINED) DURING THE FIRST YEAR OF TEACHING (CONFIRMATORY ANALYSIS)

	Condition	Means	Standard deviations	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	1.923	.277	13	-0.203	.603	-8.0%
	CREATE	1.857	.363	14			
Adjusted effect size[b]	Comparison	1.923	.277	13	-0.339	.221	-13.3%
	CREATE	1.813	.363	14			

Note. CREATE defines the group receiving the CREATE program. The *p* values are for the corresponding impact estimates in the impact model.

The Cox Index with no adjustment for effects of covariates is -.253. ($d_{Cox} = \frac{\omega\left(\left[ln\left(\frac{p_t}{1-p_t}\right) - ln\left(\frac{p_c}{1-p_c}\right)\right]\right.}{1.65}$, $\omega = [1 - 3/(4N - 9)]$)

[a] The unadjusted effect size is the regression-adjusted impact estimate from a model with a dummy variable indicating treatment status, a single dichotomous covariate indicating cohort, and random effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

We include in Appendix G the full results for the benchmark model and for two sensitivity analyses that used different approaches to selecting covariates. Those two analyses yield impact results that were similar to that of the benchmark model on the TAPS instructional strategies performance standard of -0.281 standardized effect size units (p = .396) and -0.466 standardized effect size units (p = .114)

TAPS Performance Standard 7: Positive Learning Environment

Table 20 below shows counts by response category of the Intern Keys and TAPS positive learning environment performance standard by condition and by cohort. The samples were limited to cases with non-missing baseline and outcome ratings, and who had non-missing values for several covariates used in the impact analysis, and for whom we were able to establish baseline equivalence. We observe very similar rating pattern as with the instructional strategies performance standard, with the exception that a preponderance of mid-level ratings is also more prominent on the baseline. The implications for analysis are the same as described for the instructional strategies performance standard.

TABLE 20. COUNTS FOR EACH RATING ON THE INTERN KEYS AND TAPS POSITIVE LEARNING ENVIRONMENT (*N* = 29)

	CREATE	Comparison
Counts per condition and cohort		
Cohort 1	6	10
Cohort 2	7	6
Total	13	16
Counts for each rating (cohorts 1 and 2 combined) on the Intern Keys (baseline measure)		
Level 1	0	0
Level 2	3	3
Level 3	9	12
Level 4	1	1
Counts for each rating (cohorts 1 and 2 combined) on TAPS positive learning environment (outcome)		
Level 0	0	0
Level 1	1	0
Level 2	12	14
Level 3	0	2

Note.
Intern Key ratings: Level I = Ineffective; Level II = Needs Development; Level III = Proficient; Level IV = Exemplary

TAPS ratings: Level 0 = Emerging; Level I = Developing; Level II = Proficient; Level III = Advanced

Tables G1 and G2 in Appendix G give a detailed accounting of cases and how we arrived at the final analytic samples. Across both cohorts, 15 CREATE and 16 comparison teachers had both Intern Keys ratings (baseline) and TAPS ratings (outcomes), and 13 and 16 were retained in the two conditions, respectively, for analysis. The approach to matching is described in Appendix G under "Establishing Baseline Equivalence."

The results of the impact analysis are displayed in Table 21. There was no statistically significant effect of CREATE on the positive learning environment performance standard of TAPS (p = .192). As before, we stress that the magnitude of the effect size reflects the low standard deviation associated with near-uniform ratings of teacher positive learning environment on the TAPS performance standard. Fisher's exact Test yielded results that were consistent with these. We do not reject null hypothesis of no difference between conditions in the proportions of responses across ratings categories (p = .334).

TABLE 21. IMPACT OF CREATE ON POSITIVE LEARNING ENVIRONMENT (COHORT 1 AND 2 COMBINED) DURING THEIR FIRST YEAR OF TEACHING (CONFIRMATORY ANALYSIS)

	Condition	Means	Standard deviations	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	2.125	0.342	16	-0.640	.098	-23.9%
	CREATE	1.923	0.277	13			
Adjusted effect size[b]	Comparison	2.125	0.342	16	-0.557	.192	-21.1%
	CREATE	1.950	0.277	13			

Note. CREATE defines the group receiving the CREATE program. The *p* values are for the corresponding impact estimates in the impact model.

We do not calculate the Cox Index for this outcome because $ln\left(\frac{p_c}{1-p_c}\right)$ based on the counts at posttest in Table 20 is undefined for the comparison group with $1-p_c$ in the denominator having value 0.

[a] The unadjusted effect size is the regression-adjusted impact estimate from a model with a dummy variable indicating treatment status, a single dichotomous covariate indicating cohort, and random effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

We include in Appendix G the full results for the benchmark model and for two sensitivity analyses that used different approaches to selecting covariates. Those results show similar impacts on the TAPS positive learning environment performance standard ratings of -0.611 standardized effect size units (p = .110) and -0.245 standardized effect size units (p = .488)

In this chapter, we presented results of the impact analysis—measured by the TAPS ratings—of CREATE on 1) the quality of instructional strategies, and 2) quality of the learning environment. We reiterate that the lack of impact observed may not be a fair test of the impact of CREATE, given that other considerations may have influenced the ratings. For example, raters (i.e., school principals) may have avoided giving new teachers extreme ratings, especially in high-stakes settings where ratings carry consequences. The straight-lining of ratings results in a low standard deviation, which accounts for the large magnitudes of the standardized effect sizes.

Chapter 6. Confirmatory Impacts on Student Achievement

We assessed confirmatory impacts of CREATE on **mathematics** and **ELA** achievement of students in grades 4–8, as measured by the Georgia Milestones Assessment System. Impacts on students of novice teachers in the CREATE program were evaluated at the end of their first year of teaching (that is, in their second year in the residency program), relative to students in classes of comparison teachers in their first year of teaching.

RESEARCH QUESTIONS

The impact evaluation of the CREATE teacher residency program addresses the following three *confirmatory* research questions regarding student achievement.

- What is the impact of CREATE on student **mathematics** achievement in grades 4–8, as measured by the Georgia Milestones Assessment System?
- What is the impact of CREATE on student **ELA** achievement in grades 4-8, as measured by the Georgia Milestones Assessment System?
- What is the impact of CREATE on general **(ELA and math)** achievement of students in grades 4-8, as measured by the Georgia Milestones Assessment System?

We measure impacts on student achievement for students with one year of exposure to CREATE teachers in their first year of teaching compared to students with one year of exposure to teachers in the business-as-usual group in their first year of teaching.

MEASURES

We collected student level data from the GaDOE: Georgia Milestones scores (as the outcome measure and pretest), student gender, age, grade level, race, ethnicity, special education status, and limited English proficiency status. More details about the data used in this analysis are available in the "Data Sources and Collection" section of Chapter 2.

SAMPLES

Many of the limitations described earlier that reduced the samples for analysis of impacts on teacher TAPS ratings also apply to the analysis of the student Milestones outcomes (details of the reductions in the teacher samples associated with analyses of impacts on student Milestones outcomes are in Appendix H). However, the analysis of the Milestones data included additional constraints. We matched students of CREATE teachers with those of comparison teachers on either the Math or ELA pretest within cohort and grade. We matched within cohort to ensure that the pretest scores (Milestones assessments) were collected from the same assessment administration period. Because Milestones scale scores are not vertically scaled, it is not possible to compare scores across grades. To analyze effects combined across grades, we z-transformed scores within each grade and cohort. However, where there was no representation of cases in one or the other condition within a grade and cohort, we had to discard all cases for that grade and cohort. To be included in the analysis of impact on math achievement, students had to have a math pretest score from the previous year. To be included in analysis of impact on ELA achievement, students had to have an ELA pretest score from the previous year. To be included in the analysis of impact on both outcomes combined, a student had to have an ELA pretest score, a math pretest score, or both. For the combined analysis, we pooled the matched samples obtained separately for ELA and math. With all of these factors considered, we observed a sample reduction as described in detail in Appendix H. For the grades where students

were represented in both conditions in a given cohort, we matched cases using the program *Matchit* in R (Ho, 2005; Ho et al., 2007), applying logit distances with nearest neighbor matching without replacement. The caliper, or standard deviation, of the propensity score within which comparison units were drawn was set to .25. The goal was to arrive at a sample of students of CREATE teachers who were close enough to their comparison counterparts to achieve equivalence on the pretest. If we could not find a comparison case that was sufficiently proximal to the CREATE case, we removed the CREATE case.

Additionally, to increase the sample size for the exploratory analysis, we pooled samples within grades and across cohorts. In other words, we matched CREATE and comparison students within the same grade level, regardless of cohort. This was possible because Milestones scores are horizontally scaled (allowing a direct comparison of outcomes within grades across years), which allowed us to z-transform scores within grades across cohorts. This increased the sample considerably. However, for some grade levels, outcomes for treatment cases are obtained from one cohort, while outcomes for comparison cases are obtained from the other cohort.

IMPACT MODEL

After matching students within grade and cohort (for the confirmatory analysis), we analyzed impacts on ELA, math, and across both subjects combined. The impact model used to assess impacts on math and ELA individually had the following form:

$$Y_{ij} = \beta_0 + \beta_{cohort} C_j + \beta_T T_j + \sum_{p=1}^{P} X_{p,ij} + e_{0j} + \varepsilon_{ij} \quad (3)$$

We express the z-transformed posttest score for student *i* in the class of teacher *j*, Y_{ij}, as the sum of:

- an intercept term, β_0,
- an effect of cohort, β_{cohort}, (C_j being coded 0 if belonging to Cohort 1, and 1 if belonging to Cohort 2; for analyzing confirmatory impacts on math, we removed the cohort effect, given that the analysis was based on only one cohort),
- an effect of being in treatment (T_j being coded 0 if belonging to comparison, and 1 if belonging to CREATE),
- a series of student-level covariates $X_{p,ij}$ (the covariates included the pretest, gender, ethnicity, special education status, and ELL status), and
- terms for random deviations of scores at the teacher level from the grand mean outcome conditional on covariates in the model, e_{0j}, and for random deviation of scores at the student level from the respective teacher average conditional on covariates in the model, ε_{ij}.

To evaluate impacts on math and ELA combined, we slightly adjusted the model above to account for the fact that 30 students yielded both math and ELA scores. All the students with both scores were in the CREATE condition (7th grade, Cohort 1). To address this, we included a third level to allow nesting of scores within students for the 30 cases. The model also included an additional dummy variable to indicate if the outcome score was for math or ELA.

BASELINE EQUIVALENCE

To determine baseline equivalence, we regressed the pretest against the indicator of treatment status, a dummy variable indicating cohort (where possible), and the same random effects as in the impact model. For confirmatory analyses, students of teachers in the CREATE and comparison groups were equivalent at baseline for the analysis of impact on ELA (ES = -0.06 standard deviations), on math (ES = .05 standard deviations), and on math and ELA combined (ES = -0.08 standard deviations).

For exploratory analyses, in which we allowed students of CREATE and comparison group teachers to be matched within grades and across cohorts, baseline equivalence is achieved for samples associated with analysis of impact on ELA (ES = 0.10 standard deviations) on math (ES = 0.11 standard deviations), and both subjects combined (ES = 0.12 standard deviations) (See Appendix H for details on baseline equivalence).

IMPACT FINDINGS (CONFIRMATORY)

The results of the analysis of impacts on ELA are provided in Table 22. There was no statistically significant effect of CREATE on ELA achievement (p = .454).

TABLE 22. IMPACT OF CREATE ON STUDENT ELA ACHIEVEMENT (CONFIRMATORY ANALYSIS)

	Condition	Means	Standard deviations	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	-0.18	0.94	222	9	-0.219	.280	-8.7%
	CREATE	-0.20	1.00	222	5			
Adjusted effect size[b]	Comparison	-0.18				-0.122	.454	-4.8%
	CREATE	-0.30						

Note. The *p* values are for the corresponding impact estimates in the regression model. CREATE stands for the group of students in classes of CREATE teachers.

[a] The unadjusted effect size is the impact estimate from a model with cohort, teacher, and student effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

The results of the analysis of impacts of CREATE on student math achievement are in Table 23. There was no statistically significant effect of CREATE on math achievement (p = .569).

TABLE 23. IMPACT OF CREATE ON STUDENT MATH ACHIEVEMENT (CONFIRMATORY ANALYSIS)

	Condition	Means	Standard deviations	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	-0.45	0.91	52	2	-0.133	.864	-5.2%
	CREATE	-0.87	0.77	52	1			
Adjusted effect size[b]	Comparison	-0.45				-0.175	.569	-7.1%
	CREATE	-0.60						

Note. The *p* values are for the corresponding impact estimates in the regression model. CREATE stands for the group of students in classes of CREATE teachers.

[a] The unadjusted effect size is the impact estimate from a model with cohort, teacher, and student effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

The results of the analysis of impacts of CREATE on student math and ELA achievement combined are in Table 24. There was no statistically significant effect of CREATE on both subject areas considered together ($p = .234$).

TABLE 24. IMPACT OF CREATE ON STUDENT MATH AND ELA ACHIEVEMENT COMBINED (CONFIRMATORY ANALYSIS)

	Condition	Means	Standard deviations	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	-0.23	0.94	274	11	-0.250	.272	-9.9%
	CREATE	-0.30	1.00	274	6			
Adjusted effect size[b]	Comparison	-0.23				-0.139	.234	-5.6%
	CREATE	-0.37						

Note. The *p* values are for the corresponding impact estimates in the regression model. CREATE stands for the group of students in classes of CREATE teachers.

Thirty students in the treatment condition (7th grade Cohort 1) yield both math and ELA scores (i.e., there are 518 unique student IDs). Repeated measures for these individuals were accounted for in the impact model.

[a] The unadjusted effect size is the impact estimate from a model with cohort, teacher, and student effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

IMPACT FINDINGS (EXPLORATORY)

As discussed above, we also evaluated impacts for the larger samples where we allowed students of CREATE and comparison group teachers to be matched within grades and across cohorts. These analyses are considered exploratory.

The results are exhibited as they were above for the confirmatory analyses. Impacts are summarized in Table 25, Table 26, and Table 27, with details of baseline equivalence tests in Appendix H. We see no statistically significant effects of CREATE on ELA, math, and the pooled outcomes with adjusted effect sizes of -0.067 (p = .591), 0.147 (p = .220), and -0.016 (p = .848), respectively.

TABLE 25. IMPACT OF CREATE ON STUDENT ELA ACHIEVEMENT (EXPLORATORY ANALYSIS)

	Condition	Means	Standard deviations	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	-0.25	1.00	252	14	-0.062	.748	-2.5%
	CREATE	-0.26	0.99	252	6			
Adjusted effect size[b]	Comparison	-0.25				-0.067	.591	-2.7%
	CREATE	-0.32						

Note. The *p* values are for the corresponding impact estimates in the regression model. CREATE stands for the group of students in classes of CREATE teachers.

[a] The unadjusted effect size is the impact estimate from a model with cohort, teacher, and student effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 26. IMPACT OF CREATE ON STUDENT MATH ACHIEVEMENT (EXPLORATORY ANALYSIS)

	Condition	Means	Standard deviations	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	-0.22	0.95	158	9	0.223	.400	8.8%
	CREATE	-0.23	0.90	158	6			
Adjusted effect size[b]	Comparison	-0.22				0.147	.220	5.8%
	CREATE	-0.08						

Note. The *p* values are for the corresponding impact estimates in the regression model. CREATE stands for the group of students in classes of CREATE teachers.

[a] The unadjusted effect size is the impact estimate from a model with cohort, teacher, and student effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

TABLE 27. IMPACT OF CREATE ON STUDENT MATH AND ELA ACHIEVEMENT COMBINED (EXPLORATORY ANALYSIS)

	Condition	Means	Standard deviations	No. of students	No. of teachers	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	Comparison	-0.24	0.98	410	20	0.043	.831	1.7%
	CREATE	-0.25	0.96	410	8			
Adjusted effect size[b]	Comparison	-0.18				-0.016	.848	-0.6%
	CREATE	-0.21						

Note. The *p* values are for the corresponding impact estimates in the regression model. CREATE stands for the group of students in classes of CREATE teachers.

95 students in the treatment condition and 13 students in the comparison group yield both math and ELA pretest scores (i.e., there are 712 unique student ID's). Repeated measures for these individuals were accounted for in the model used to assess impact.

[a] The unadjusted effect size is the impact estimate from a model with cohort, teacher, and student effects, divided by the pooled standard deviation of the outcome variable.

[b] The adjusted effect size estimate is the point estimate for impact from the benchmark model divided by the pooled standard deviation of the outcome variable.

Chapter 7. Exploratory Impacts on Early Career Teaching Trajectories and Retention

In this chapter, we address exploratory impacts of CREATE on retention. More specifically, we examine impacts on early career teachers' three-year trajectory, starting with graduation from GSU CEHD, and into the first and second year of teaching.

RESEARCH QUESTIONS

The impact evaluation of CREATE addresses the following two *exploratory* research questions regarding early career teaching trajectories.

- What is the impact of CREATE on completion of the teacher preparation program at GSU CEHD and teacher retention into the first and second year of teaching for the overall sample and for Black educators?
- Is there a differential impact for Black and non-Black educators?

DEFINITIONS OF KEY TERMS

We define the outcome of each step in the three-year trajectory in the following ways.

Graduation from GSU College of Education and Human Development in Year 1. Graduation from GSU CEHD in either the Early Childhood and Elementary Education program or the Middle and Secondary Education program by the summer of their expected graduation year.

Teaching in Year 2 (first year of teaching). In Year 2, a teacher is considered to be teaching if they are employed in a teaching position as a teacher-of-record, an associate teacher, a paraprofessional, a support teacher, a long-term substitute teacher, or an online teacher in a K-12 public school in Georgia.

Teaching in Year 3 (second year of teaching). In Year 3, a teacher is considered to be teaching if they are employed in a teaching position as a teacher-of-record, an associate teacher, or an online teacher in a K–12 public school in Georgia.

MEASURES

Early Career Three-Year Trajectory

Determining Teacher Status

We rely on a variety of sources to determine status for each study participant at the three time points. For graduation from GSU CEHD, we rely on data from our participant tracker[9], participant surveys, data provided to the research team by GSU or the CREATE program team, and teacher certification records from the Georgia Professional Standards Commission (Georgia Professional Standards Commission, 2014). For teaching in Year 2 and teaching in Year 3, we triangulate data received on teacher surveys, data received from the CREATE program team, data from GADOE, and teaching records from Open Georgia: Transparency in Government travel and salary database (Open Georgia, 2008).

[9] The participant tracker is a database of all study participants and their contextual information, including demographic characteristics, teacher preparation program, practicum and teaching placements (e.g., district, school, grade), data collection completion, and notes from any communication with or about the participant with GSU CEHD and the CREATE program team.

Each participant has a record indicating their early career trajectory for the first three years, with the first year covering graduation from GSU CEHD, the second indicating teaching status for the first year after graduation, and the third indicating teaching status for the second year after graduation. For each of the three years, we code outcomes for participants with 0 (not graduated or not teaching), 1 (graduated or teaching), or 2 (unknown status).[10] For example, a participant who graduated and was a teacher of record in each of the following two years has a retention outcome across the three years of 1, 1, 1. A participant who graduated but who did not teach in the two subsequent years, has a record of 1, 0, 0. A participant who graduated, taught the next year, and for whom we cannot verify status in the second year has outcomes coded as 1, 1, 2. Appendix I includes additional details on how we coded the records for the three-year teaching trajectory.

The Churn Factor

Not all teachers start teaching right after graduation. In some cases, positions may not be available immediately or participants may choose to take a year out of the classroom for personal reasons. That is, there may be a period of "churn" during which teachers establish their career trajectories. For example, a teacher may graduate, spend a year applying for work, and then start teaching the second year after graduation; thus they receive an outcomes record of 1, 0, 1. Alternatively (and rarely), a teacher may not graduate on time, finish their program in the second year and start teaching in the third year, and thus receives a record of 0, 2, 1. In this case, we do not know their status in the second year, however, we infer that they completed their program and graduated that year in order for them to become a teacher of record in the third year.

We take two approaches in producing results to facilitate interpretation of teaching status. For descriptive analyses, we counted individuals based on their status in a given year, independent of their graduation or teaching status the prior year(s). That is, we count the number of cases—as described in the previous paragraph—independently for each year over the three-year early career trajectory. For analysis of longitudinal trends (for which we use discrete-time survival analysis, as we describe further below), we base retention status on on-time teaching. *On-time* graduating and transitioning to teaching is defined as "a teacher graduates the year they enroll and becomes a teacher of record in the two consecutive years after graduating from GSU CEHD." In that case, a person has a retention record of 1, 1, 1. If a teacher graduates but does not become a teacher of record until a full year later (1, 0, 1), their status is recoded to 1, 0, 0. In this case, they start their career but the transition to teaching is lagged. Similarly, the hypothetical case above with a record of 0, 2, 1 would be recoded to 0, 0, 0. We apply the rule that once a participant is coded 0, then all following timepoints are recoded to 0.

Both types of outcomes are important for informing policy. The first indicates whether a teacher, at some point within the three-year trajectory, graduates and enters teaching. The second addresses whether a teacher transitions into teaching on time, indicating a faster and higher return on investment in a teaching career. In Appendix I, we show the frequency of counts in each category of the three-year trajectories, after recoding teaching status in the ways described here.

Additional Variables Used in The Analysis

We employed additional teacher-level variables in the analysis: an indicator of whether a teacher is a Black educator or not, whether a teacher belongs to the first or second study cohort, levels of confidence in general teaching skills,

[10] Additional information about possible scenarios that resulted in an unknown status, the follow-up process taken to track down status, and frequency counts of participants within each category of unknown status is provided in Appendix I.

motivations for entering the teaching profession, levels of math anxiety, and self-reported postsecondary GPA. We collected data for such variables through the baseline survey. Appendix F includes the description of baseline survey data.

MATCHING AND RESULTING SAMPLES

Because the study is a quasi-experiment, we take steps to ensure that we are comparing similar cases in both conditions, to rule out the potential effects of confounds that would bias the impact estimates. To some extent, the fact that all participants joined the teaching program at GSU CHED assured similarity by suggesting similar motivations, interests, and geographic residency while enrolled. However, additional selection effects may be related to the motivations of CREATE participants to enter CREATE. Therefore, in the survival analysis described below, we analyze impacts after matching CREATE and comparison cases within each cohort on a series of baseline covariates to achieve equivalence on those variables.[11] The covariates consisted of responses obtained at baseline about teachers' confidence in general teaching skills, motivations for entering the teaching profession, math anxiety, and self-reported postsecondary GPA. We conducted matching for the full sample of cases available (Black educators and non-Black educators combined) and for Black educators only for the analysis limited to just the subgroup.

For the full sample, we removed missing covariate data for 2 CREATE teachers and 11 comparison teachers from analysis, leaving 38 and 83 cases in the two conditions, respectively. The remaining sample was balanced in terms of covariates without any matching (the difference between CREATE and comparison groups for each of the four baseline covariates was less than 0.25 standard deviations).

The sample of 53 Black educators was not missing any covariate data. However, for this sample, baseline equivalence was not achieved to start with. We matched cases using the program "Matchit" in R (Ho, 2005; Ho et al., 2007), applying logit distances with nearest neighbor matching without replacement. The caliper, or standard deviation of the propensity score within which we drew comparison units, was set to .50. The goal was to arrive at a sample of CREATE participants who were close enough to their comparison counterparts to achieve equivalence. If we could not find a comparison case that was sufficiently proximal to the treatment case, we removed the treatment case. Following matching, we retained 19 of 22 Black CREATE teachers, and 19 of 31 Black comparison teachers for analysis.

ANALYSIS

Descriptives

For the descriptive analysis, we examined the proportions of teachers who graduated, who were teachers of record in Year 2 (first year of teaching), and who were teachers of record in Year 3 (second year of teaching), using the full sample of 40 CREATE teachers and 94 comparison teachers across both cohorts prior to matching. For each year, using cases not lost to follow-up, we calculated the proportions who graduated or who were retained. We used three approaches to test the

[11] While CREATE and comparison teachers had matched experience (they were all novices), we could not match them in terms of the baseline characteristics of the students in the classes of their cooperating or mentor teachers during the year of residency due to the unavailability of those data. This is a limitation of the current work.

difference between CREATE and comparison cases in these proportions.[12] We present these results for the full sample and for the subsample of Black educators in Table 28 and Table 29 below. We summarize the impact using the Cox index, the effect size measure of choice for dichotomous outcomes.

We underscore that the descriptive analysis is intended to show data available in each category of outcomes for each year and condition. Therefore, cases counted as not graduated or not teaching in a year (with outcomes marked as 0) could still receive any of the possible outcomes (0, 1, or 2) in the subsequent year(s). This is different from the survival analysis described below, where individuals, once marked as 0 for a certain outcome, maintain this same outcome in subsequent years.

Survival Analysis

In addition to calculating descriptives, we assess the impact of CREATE on on-time graduation and teaching in their first two years. One of the benefits of using a model-based approach is that it enables us to address right-censored data; that is, we can include everyone in analysis and address the problem that we lose information about the retention status of individuals in certain time intervals. In this study, this problem is especially prevalent in the comparison group between the first and second years of teaching. Survival analysis (which is labeled as such because the methodology comes from research on factors related to survival in the health sciences) provides a means to infer the status of individuals over three years, taking into account that individuals are lost to follow-up over that time.

In this work, we use "discrete-time survival analysis" (Singer & Willett, 1993) to evaluate the impact of CREATE on the three-year early career trajectory (graduation from GSU CEHD in Year 1, teaching in Year 2, and teaching in Year 3). The method produces maximum-likelihood estimates of key parameters in the model. The general form of the impact model is as follows.

$$log_e\left(\frac{h_{ij}}{1-h_{ij}}\right) = \alpha_1 time_{1ij} + \alpha_2 time_{2ij} + \alpha_3 time_{3ij} + \beta_0 treatment_{ij} + \beta_1 Cohort_{ij} + [\beta_2 BlackEd_{ij} + \sum_{k=1}^{N} \gamma_k X_{ij}] \quad (4)$$

Here, i indexes the individual, and j indexes the time period (time = 1, 2, or 3). The hazard, h_{ij}, is the probability that individual i experiences the event (leaving the residency program or teaching) in period j, given that that person has not left teaching in any prior period. The log odds of the hazard is expressed as a linear function of time ($time_{1ij}$, $time_{2ij}, time_{3ij}$, each coded 0 or 1 depending on the period in which the outcome is observed), treatment status ($treatment_{ij}$, coded 0 for comparison and 1 for treatment), cohort ($cohort_{ij}$, coded 0 for cohort 1, and 1 for cohort 2) whether a teacher is a Black Educator ($BlackEd_{ij}$ coded 0 for non-Black educator and 1 for Black educator), and a series of non-time-varying covariates (X_{ij}).

[12] The three approaches were (1) a standard test of the difference in proportions, assuming that outcomes for individuals are binomially distributed, (2) a logistic regression with the log odds of retention regressed against an indicator of treatment status and an indicator of cohort, and (3) Fisher's exact test. The last method is important because the high retention rate among CREATE teachers results in low counts within certain cells in the cross-tabulation between condition and retention status. Under these conditions, parametric tests can be unstable, and exact methods are recommended.

From the estimates, we calculate the fitted hazards function $\widehat{h_{ij}}$, which is the estimated probability that an event (leaving teaching) occurs in a given time period, given that it has not occurred up to that timepoint. We also report the *survivor probabilities*, which are the probabilities that a teacher stays retained in a given time interval, given that they have been retained up to that time interval (i.e., it is $1-\widehat{h_1}$ through interval 1, $\left(1-\widehat{h_1}\right)\left(1-\widehat{h_2}\right)$ through interval 2, and $(1-\widehat{h_1})(1-\widehat{h_2})(1-\widehat{h_3})$ through interval 3).

The resulting sample survival function, which is the set of the survivor probabilities over time, is the most intuitive way to interpret the impact of CREATE on the three-year early career trajectory. It tells us the proportion of teachers (based on the number of teachers in the initial sample, i.e., at the beginning of interval 1) expected to remain in the early career trajectory by the end of each interval. At a given timepoint, the value of the function is the proportion of teachers still teaching at that time point. For example, the function allows us to estimate by when 25% of teachers have left teaching (or equivalently, at what point 75% of teachers still remain in teaching) separately for CREATE and comparison groups.[13]

We use the following models with the full matched sample to assess impacts of CREATE on rates of retention, as well as differential impacts between Black and non-Black educators in rates of retention.

1. Model 0: the base model with the log odds of the hazard regressed against time covariates and a variable indicating cohort membership.

$$log_e\left(\frac{h_{ij}}{1-h_{ij}}\right) = \alpha_1 time_{1ij} + \alpha_2 time_{2ij} + \alpha_3 time_{3ij} + \beta_1 Cohort_{ij} \quad (5)$$

2. Model 1: like Model 0, but also includes a term indicating if a teacher belongs to CREATE or the comparison group. The results from Model 1 are compared to those from Model 0 to determine if there is an overall impact of CREATE.

$$log_e\left(\frac{h_{ij}}{1-h_{ij}}\right) = \alpha_1 time_{1ij} + \alpha_2 time_{2ij} + \alpha_3 time_{3ij} + \beta_0 treatment_{ij} + \beta_1 Cohort_{ij} \quad (6)$$

[13] Note that the methodology assumes that censoring of observations is unrelated to event occurrence, known as *independent censoring* (Singer & Willett, 1993), which means the risk of an event (lost from teaching) for a given subgroup in a given year is the same for everyone, regardless of whether in the course of that year, a person is lost to follow-up (i.e., resulting in a censored observation) or not.

In the survival analysis of this work, we should keep in mind the potential for instability of estimates that may result from there being small numbers of cases in some of the survival categories for certain subgroups.

3. Model 2: like Model 1, but also includes the main effect of retention on being a Black educator.

$$log_e\left(\frac{h_{ij}}{1-h_{ij}}\right) = \alpha_1 time_{1ij} + \alpha_2 time_{2ij} + \alpha_3 time_{3ij} + \beta_0 treatment_{ij} + \beta_1 Cohort_{ij} + \beta_1 BlackEd_{ij} \quad (7)$$

4. Model 3: like Model 2 but also includes a term for the interaction between the variable indicating whether a teacher is a Black educator and the variable indicating treatment status. The results from Model 3 are compared to those from Model 2 to determine if the impact of CREATE on retention differs between Black and non-Black educators.

$$log_e\left(\frac{h_{ij}}{1-h_{ij}}\right) = \alpha_1 time_{1ij} + \alpha_2 time_{2ij} + \alpha_3 time_{3ij} + \beta_0 treatment_{ij} + \beta_1 Cohort_{ij} + \beta_1 BlackEd_{ij} + \beta_2 BlackEd_{ij} \times treatment_{ij} \quad (8)$$

For each of these models, we report the estimated regression coefficients and the deviance statistic (-2 × the log likelihood). If the difference in the deviance (Model 1 versus Model 0 for assessing impact of CREATE, or Model 3 versus Model 2 for assessing the differential impact of CREATE) reaches statistical significance, then we can conclude the added effect is statistically significant.[14] We also report the fitted hazard and survivor functions.

We evaluated four additional models (Models 4 – 7), which correspond to Models 0 – 3, but also include the four baseline covariates used to match cases. Our main findings are based on results from Models 4 – 7.

Additionally, for analyzing impacts on just the sample of Black educators, we used similar models to those above. Specifically, we compared Model 1 to Model 0, first without baseline covariates, and then after including those covariates.

RESULTS

Descriptives

The sample for the descriptive analysis consists of 40 CREATE residents and 94 comparison participants across two cohorts who were eligible for the study, agreed to participate, and represented the baseline sample for the analysis of retention outcomes. Table 28 shows the sample, tracked across three years, with outcomes pooled across cohorts. We observe a difference in on-time graduation rates favoring CREATE, with 39/40 (98%) of CREATE residents and 80/94 (85%) of comparison participants graduating. The difference of 13% in proportion graduating is statistically significant based on

[14] The difference in the deviance statistic between models being compared (Model 1 versus Model 0 for assessing impact of CREATE, or Model 3 versus Model 2 for assessing the differential impact of CREATE) follows a Chi-squared distribution with degrees of freedom equal to the difference between models compared in the number of parameters being estimated.

Fisher's exact test (p = .039). Parametric tests of the difference in proportions based on logistic regression are consistent with this result.

For teaching in Year 2 (first year of teaching), we also observe a difference that favors CREATE in the proportion of participants teaching, with 38/40 (95%) of CREATE residents and 62/86 (72%) comparison participants teaching (with 8 comparison cases lost to follow-up). The difference (23%) is statistically significant based on Fisher's exact test (p = .001). Parametric tests of the difference in proportions based on logistic regression are consistent with this result.

For teaching in Year 3 (second year of teaching), there was no statistically significant difference between the CREATE and the comparison group in the proportion teaching. However, we observe that in Year 3, many of the comparison group are lost to follow-up. This is because either we were unable to collect data from the district or participants opted out of the study, and no other data sources were able to verify teaching status. In CREATE, 1 teacher out of the original 40 was lost to follow-up. Of the 39 remaining, 34 were retained through the second year of teaching, and 5 are no longer teaching. In the comparison group, 24 of the original 94 were lost to follow-up, and of the 70 remaining, 64 are retained through the second year of teaching (6 are no longer teaching). Given this large number of comparison cases for whom we do not have teaching status during the second year of teaching, the non-statistically significant difference in proportion teaching is not reliable. For example, we expect that many of the 24 comparison cases not retained in teaching in the first year after graduation will also not be teachers of record the following year; however, we cannot verify their status in Year 2 for the reasons mentioned above (i.e., opted out of study or unable to secure district data).[15] The survival analysis results below (in the next section), which figure in the cases lost to follow-up, address this limitation in the outcomes.

[15] See Appendix I for additional details on cases that were lost to follow-up.

TABLE 28. NUMBERS RETAINED FOR THE FULL SAMPLE

	CREATE				Comparison group				Impact		
	Sample size	Number missing [a]	Number graduated or teaching	Percentage	Sample size	Number missing [a]	Number graduated or teaching	Percentage	CREATE - comparison difference	*p* value [b]	Cox Index
Graduated from GSU CEHD in Year 1	40	0	39	97.5%	94	0	80	85.1%	12.4%	.039	1.16
Remained in teaching in Year 2 (first year of teaching)	40	0	38	95.0%	86	8	62	72.1%	22.9%	.001	1.20
Remained in teaching in Year 3 (second year of teaching)	39	1	34	87.2%	70	24	64	91.4%	-4.2%	.518	-0.27

[a] Teaching status is unknown because it could not be verified through the various sources of data available to the research team. See Appendix I for additional details on coding of teaching status.

[b] *p* values reported are from Fisher's exact test, due to the small sample sizes. Alternative approaches to the statistical test (Chi-square test of difference in proportions and logistic regression) yielded similar *p* values.

We also examine retention results with the sample limited to Black educators. We observe a similar pattern to that observed for the full sample. On-time retention is remarkable among Black educators in CREATE, with all 22 cases graduating, and 21/22 (96%) of cases retained through teaching in Year 2 (first year of teaching) and in Year 3 (second year of teaching), and with none lost to follow-up. In the comparison group 24 of 31 residents (77%) graduated. For teaching in Year 2 (first year of teaching), we lost two cases to follow-up. Of the remainder, 21 of 29 (72%) stayed in teaching. In teaching in Year 3 (second year of teaching), nine are lost to follow-up. Of the remainder, 21 of 22 (96%) stayed in teaching. We observed a statistically significant difference favoring CREATE in graduation rate ($p = .017$) and a marginally significant difference in proportion teaching in Year 2 ($p = .060$), with results based on Fisher's exact test. The difference between conditions in retention among teachers not lost to follow-up through teaching in Year 3 is not statistically significant; however, as with the full sample, in the comparison group, many of those marked as not teaching in the first year (Year 2) are lost to follow-up in Year 3, likely resulting in an underestimation of the impact of CREATE. Next, we address the results of the survival analysis.

TABLE 29. NUMBERS RETAINED FOR THE BLACK EDUCATORS SAMPLE

	CREATE				Comparison group				Impact		
	Sample size	Number missing [a]	Number graduated or teaching	Percentage	Sample size	Number missing [a]	Number graduated or teaching	Percentage	CREATE - comparison difference	*p* value [b]	Cox Index
Graduated from GSU CEHD in Year 1	22	0	22	100%	31	0	24	77.4%	22.6%	.033	n/a [c]
Remained in teaching in Year 2 (first year of teaching)	22	0	21	95.5%	29	2	21	72.4%	23.1%	.060	1.25
Remained in teaching in Year 3 (second year of teaching)	22	0	21	95.5%	22	9	21	95.5%	0%	1.000	0.00

[a] Teaching status is unknown because it could not be verified through the various sources of data available to the research team. See Appendix I for additional details on coding of teaching status.

[b] *p* values reported are from Fisher's exact test, due to the small sample sizes. Alternative approaches to the statistical test (chi-square test of difference in proportions and logistic regression) yielded similar *p* values.

[c] The Cox index is undefined in this instance with odds of success in treatment being (1 / (1 – 1)).

Main Impact Findings

Results of Matching

As noted earlier, we conducted matching twice using four baseline covariates. First, we matched cases across conditions for the full sample available. Second, we matched cases among Black educators only. To be included in the matching procedure, participants had to have no missing values for any of the covariates on which matching was conducted. The sample sizes prior to and after matching in each condition are dispalyed in Table 30 for the full sample and in Table 31 for the sample limited to Black educators.

We observe in Table 30 that we lost some teachers from the analysis because they had missing values for some of the covariates. However, after limiting the sample to teachers with non-missing values for the covariates, we retained the full sample in the analysis (i.e., baseline equivalence was achieved for the sample as a whole, with the average difference between conditions being less than .25 SD for each baseline covariate). We observe in Table 31 that when limiting the analysis to the sample of Black educators only, all teachers had complete covariate data; however, the samples were non-equivalent to start, and matching of cases across conditions and within cohort led to a reduction in the sample.

TABLE 30. SAMPLE SIZES BEFORE AND AFTER MATCHING FOR THE FULL SAMPLE

	Full sample	Limited to teachers with non-missing baseline covariates	After matching[a]
CREATE Cohort 1	19	19	19
CREATE Cohort 2	21	19	19
Comparison Cohort 1	56	50	50
Comparison Cohort 2	38	33	33

[a] After limiting the sample to participants with non-missing values for the covariates, we were able to retain the full sample in analysis because baseline equivalence was achieved for the sample as a whole.

TABLE 31. SAMPLE SIZES BEFORE AND AFTER MATCHING FOR THE SUBSAMPLE OF BLACK EDUCATORS

	Full sample	Limited to teachers with non-missing bsaeline covariates	After matching
CREATE Cohort 1	12	12	10
CREATE Cohort 2	10	10	9
Comparison Cohort 1	23	23	13
Comparison Cohort 2	8	8	6

Results of tests of baseline equivalence for each of the covariates—for the both the full sample (N = 121) and the sample of Black educators (N = 38)—used in analysis are provided in Appendix J. In all cases, baseline equivalence was achieved

(with standardized differences < .25 standard deviations), with the requirement that the baseline variables on which equivalence was assessed are included in the impact model.

Results of Survival Analysis Based on The Full Sample

Table 32 shows the main results of the survival analysis across eight models. Model 0–7 were described earlier under the Analysis section.

We observe a negative and statistically significant impact of CREATE on the log odds of the hazard probability of undisrupted retention in a three-year early career trajectory (spanning graduation from GSU CEHD and first and second years of teaching) for the CREATE group, compared to the comparison group (p = .038) (Model 5). We also observe that the favorable impact is driven largely by higher continuous retention among Black educators in CREATE, relative to those in the comparison group (p = .021) (Model 7). The results indicate a reduced probability of dropout from the three-year early career trajectory in CREATE, relative to the comparison group, and that the favorable impact is largely due to the high retention among Black educators in CREATE.

TABLE 32. RESULTS OF THE SURVIVAL ANAYSIS FOR THE FULL SAMPLE

Predictor	Model 0 Parameter Estimate (SE)	Model 1 Parameter Estimate (SE)	Model 2 Parameter Estimate (SE)	Model 3 Parameter Estimate (SE)	Model 4 Parameter Estimate (SE)	Model 5 Parameter Estimate (SE)	Model 6 Parameter Estimate (SE)	Model 7 Parameter Estimate (SE)
Time 1	-2.260 (0.31)	-2.048 (0.32)	-1.974 (0.35)	-2.143 (0.37)	1.367 (2.59)	1.412 (2.65)	1.419 (2.72)	1.261 (2.72)
Time 2	-1.869 (0.28)	-1.630 (0.30)	-1.560 (0.33)	-1.708 (0.34)	1.794 (2.61)	1.875 (2.66)	1.882 (2.73)	1.750 (2.74)
Time 3	-2.909 (0.46)	-2.645 (0.48)	-2.569 (0.49)	-2.745 (0.51)	0.760 (2.63)	0.877 (2.68)	0.883 (2.75)	0.721 (2.76)
Belongs to Cohort 1	0.231 (0.20)	0.210 (0.20)	0.233 (0.20)	0.208 (0.20)	0.231 (0.20)	0.224 (0.21)	0.224 (0.21)	0.224 (0.21)
Belongs to CREATE		-0.857 (0.47)	-0.807 (0.48)	0.028 (0.55)		-0.930 (0.48)	-0.928 (0.50)	-0.060 (0.57)
Is a Black educator			-0.222 (0.40)	0.226 (0.44)			-0.005 (0.51)	0.530 (0.55)
Added value impact of CREATE for Black educators				-2.346 (1.20)				-2.457 (1.21)
Current GPA					-0.039 (0.43)	0.002 (0.43)	-0.002 (0.51)	0.092 (0.51)
Confidence in teaching					-0.754 (0.36)	-0.819 (0.37)	-0.819 (0.38)	-0.877 (0.40)
Motivation to teach					0.058 (0.35)	0.117 (0.35)	0.117 (0.35)	0.090 (0.35)
Level of math anxiety					-0.314 (0.20)	-0.301 (0.21)	-0.301 (0.21)	-0.309 (0.21)
-2 × Log Likelihood	202.793	199.031	198.725	193.754	196.554	192.255	192.255	186.903
Change in -2 × Log Likelihood (*df*)		3.762 (1)	0.306 (1)	4.971 (1)	6.239 (4)	4.299 (1)	0 (1)	5.352 (1)
***p* value**		.052	.580	.026	.182	.038	1.000	.021

Note. The subgroup sample sizes for each of the analyses were 38 CREATE cases (19 from each cohort) and 83 comparison cases (50 from Cohort 1 and 33 from Cohort 2).

Model comparisons are between Models 1 and 0, 2 and 1, 3 and 2, 4 and 0, 5 and 4, 6 and 5 and 7 and 6.

The intuition for the results is most easily captured through the survival percentages for the CREATE and comparison groups for the full sample, the sample of non-Black educators, and the sample of Black educators in Tables 33–35. The tables display percentages of teachers who left their teaching career trajectory within each year, and the percentages who remained in each time interval. The percentages of teachers remaining are also displayed graphically in Figures 12–14[16].

The percentages of teachers in CREATE, who maintain an uninterrupted trajectory of (1) graduating from GSU CEHD, (2) teaching in their first year after graduating, and (3) teaching in their second year after graduating are 95%, 87% and 85%, respectively. In the matched comparison group, the corresponding values are 88%, 73%, and 68% (see Table 33 and Figure 12). This demonstrates the average positive impact of CREATE. Among non-Black teachers, the values in the CREATE condition are 90%, 75% and 71%, respectively, and in the comparison condition, they are nearly identical (89%, 75% and 71%) (see Table 34 and Figure 13). However, among Black teachers, the difference between these trajectories is substantive: for CREATE residents, the values are 99%, 96% and 96%, compared to 86%, 69%, and 63%, respectively, for the comparison condition (see Table 35 and Figure 14). This shows that the average positive impact of CREATE on retention is driven by the differentially greater impact of CREATE on Black educators compared to non-Black educators.

TABLE 33. THE PERCENTAGES OF TEACHERS LEAVING OR REMAINING ON THEIR CAREER TRAJECTORIES FOR THE FULL SAMPLE

	CREATE		Comparison	
	% left during year	% remaining (survival percentages)	% left during year	% remaining (survival percentages)
Year 0: At start of GSU CEHD		100%		100%
Year 1: Year of GSU CEHD	5.2%	94.8%	12.1%	87.9%
Year 2: First year teaching	7.8%	87.4%	17.1%	72.9%
Year 3: Second year teaching	3.2%	84.6%	6.8%	68.0%

[16] To calculate the percentages, we generated fitted values for the log odds of hazard probabilities for each timepoint, converted them to hazard probabilities, and calculated the means for each subgroup of interest. All results are based on Model 7 in Table 32.

TABLE 34. THE PERCENTAGES OF TEACHERS LEAVING OR REMAINING ON THEIR CAREER TRAJECTORIES FOR THE NON-BLACK EDUCATOR SAMPLE

	CREATE		Comparison	
	% left during year	% remaining (survival percentages)	% left during year	% remaining (survival percentages)
Year 0: At start of GSU CEHD		100%		100%
Year 1: Year of GSU CEHD	10.4%	89.6%	10.8%	89.2%
Year 2: First year teaching	15.8%	75.4%	15.6%	75.4%
Year 3: Second year teaching	6.4%	70.6%	5.82%	71.0%

TABLE 35. THE PERCENTAGES OF TEACHERS LEAVING OR REMAINING ON THEIR CAREER TRAJECTORIES FOR THE BLACK EDUCATOR SAMPLE

	CREATE		Comparison	
	% left during year	% remaining (survival percentages)	% left during year	% remaining (survival percentages)
Year 0: At start of GSU CEHD		100%		100%
Year 1: Year of GSU CEHD	1.4%	98.6%	14.3%	85.7%
Year 2: First year teaching	2.3%	96.3%	20.0%	68.5%
Year 3: Second year teaching	0.8%	95.5%	8.4%	62.8%

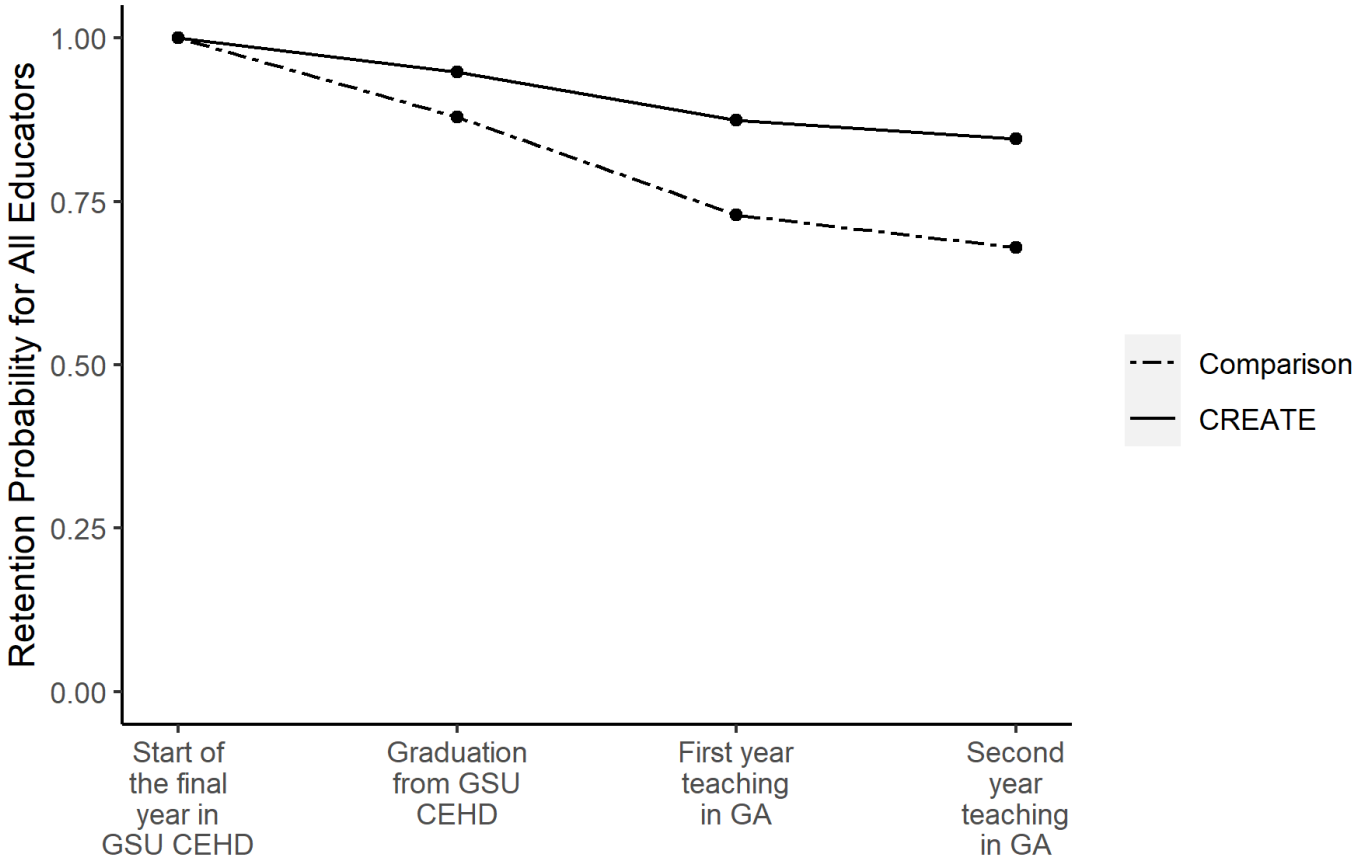

FIGURE 12. THE PERCENTAGES OF TEACHERS WHO REMAINED ON AN UNINTERRUPTED CAREER TRAJECTORY IN CREATE AND COMPARISON BY COHORT (FULL SAMPLE)

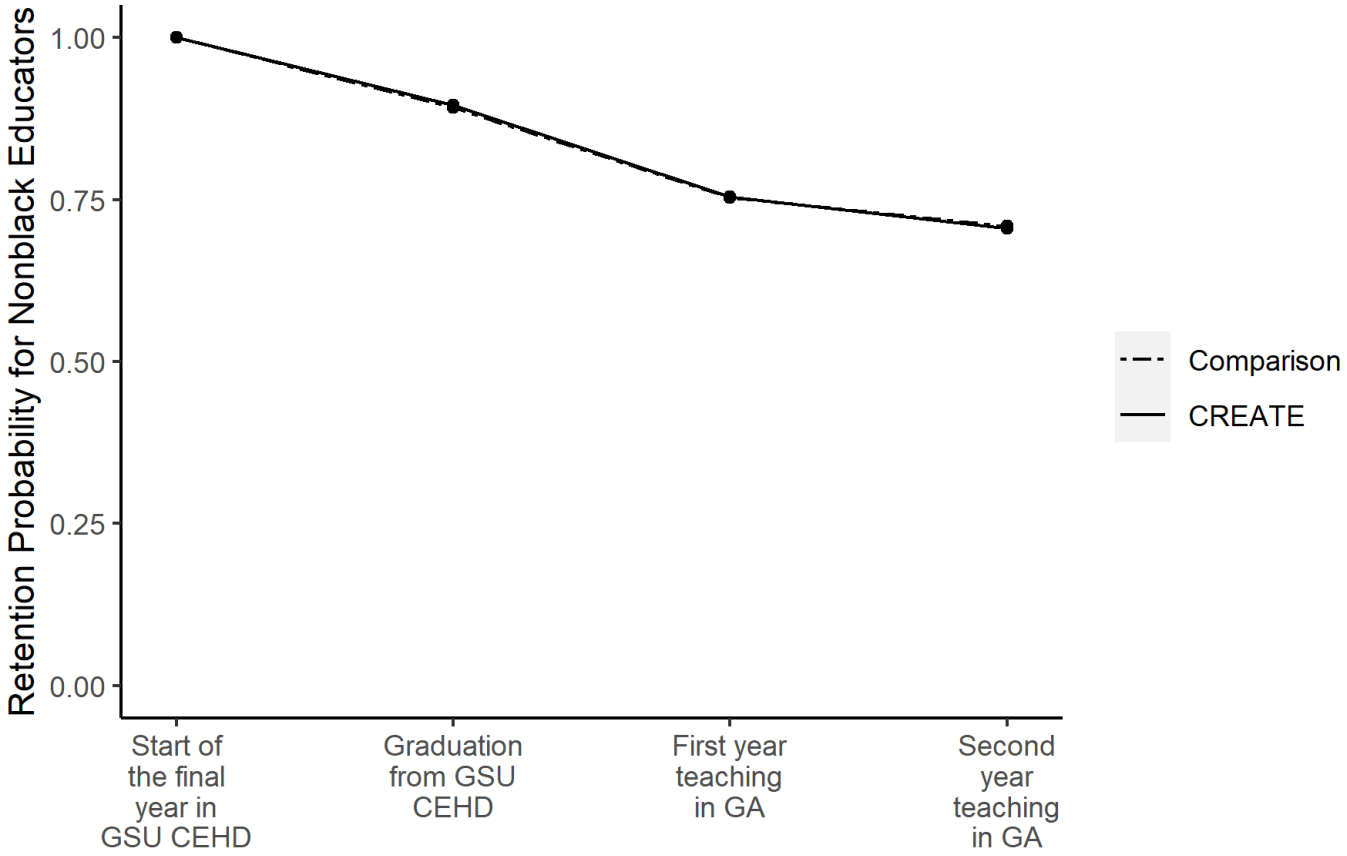

FIGURE 13. THE PERCENTAGES OF TEACHERS WHO REMAINED ON AN UNINTERRUPTED CAREER TRAJECTORY IN CREATE AND COMPARISON BY COHORT (SAMPLE OF NON-BLACK EDUCATORS)

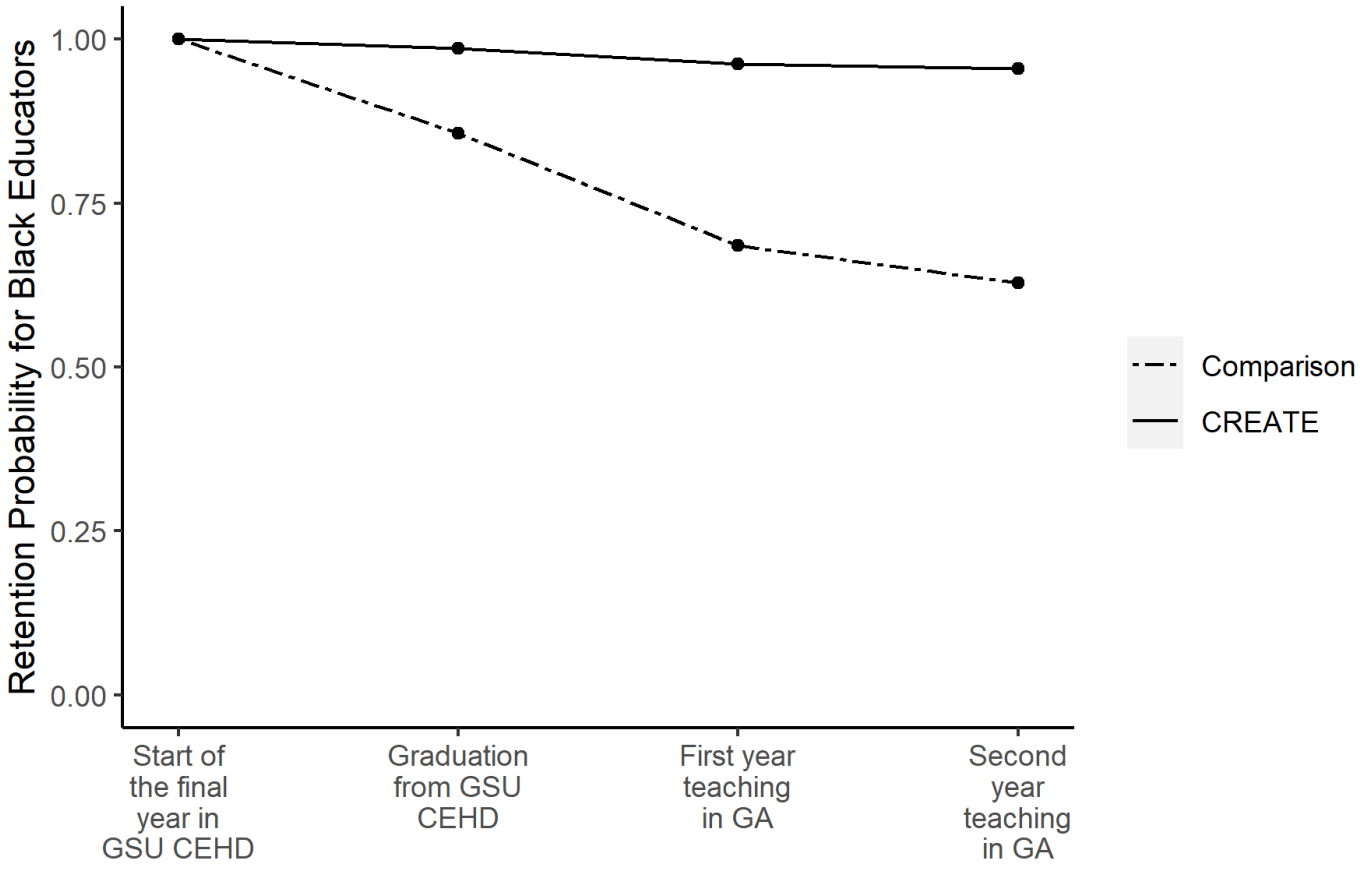

FIGURE 14. THE PERCENTAGES OF TEACHERS WHO REMAINED ON AN UNINTERRUPTED CAREER TRAJECTORY IN CREATE AND COMPARISON BY COHORT (SAMPLE OF BLACK EDUCATORS)

Results of Survival Analysis Based on The Black Educators Sample Only

Additionally, we examined impacts for Black educators only. The rationale for this is that when matching samples within this subgroup, some cases were removed (mostly from the comparison group) to achieve equivalent samples. (In contrast, for the full sample, no cases were removed in the matching process.) The question of interest is whether the benefits of CREATE for Black educators that we observed with the full sample are sustained when we limit our analysis to just the matched sample of Black educators.

Table 36 shows the main results of the survival analysis. The first two models (Model 0 and Model 1) do not incude covariates, while the latter two models (Model 2 and 3) do include covariates. Because this sample consists of Black

educators only, we removed the variable indicating if a teacher is a Black educator and the interaction of that variable with the one indicating membership in CREATE.[17]

We observe a negative and statistically significant impact of CREATE on the log odds of the hazard probability of undisrupted retention over a three-year time period (spanning graduation from GSU CEHD and first and second year of teaching) for the CREATE group, compared to the comparison group for both Model 1 (without covariate adjustment), with an impact estimate of -2.49 (p = <.01), and Model 3 (with covariate adjustments), with an impact estimate of -2.55 (p < .01). The results indicate a reduced probability of dropout from the career trajectory over the three-year time period in CREATE, relative to the comparison group. The impact estimates are similar across the two models. They are also similar to results based on the full sample reported in Table 32, specifically, the estimates of added value impact of CREATE on retention for Black educators (-2.35, p = .026) for Model 3 (without covariates), and -2.46 (p = .021) for Model 7 (with covariates). This suggests that the effect observed with full sample estimates (reported in Table 32) is robust due to the reduction of the sample of Black educators through matching and then limiting analysis to this matched sample. It also reaffirms that there is a positive impact among Black educators and that the effect of CREATE on retention in the three-year early career trajectory for the full sample (i.e., combining samples of Black and non-Black educators) is driven by impact on Black educators.

[17] We encourage the reader to interpret the results presented in this section with caution. The estimation procedure issued warnings about model convergence that is likely due to very few Black teachers in CREATE leaving the teaching trajectory. However, results are consistent with those in Table 32, where there were no estimation issues with the full sample for any of the eight models.

TABLE 36. RESULTS OF THE SURVIVAL ANAYSIS LIMITED TO THE MATCHED SAMPLE OF BLACK EDUCATORS

Predictor	Model 0 Parameter Estimate (SE)	Model 1 Parameter Estimate (SE)	Model 2 Parameter Estimate (SE)	Model 3 Parameter Estimate (SE)
Time 1	-1.857 (0.48)	-1.066 (0.54)	4.758 (6.44)	9.16 (8.96)
Time 2	-2.274 (0.61)	-1.349 (0.67)	4.383 (6.48)	8.99 (9.04)
Time 3	-14.16 (221)	-13.14 (206)	-7.481 (216)	-2.85 (203)
Belongs to Cohort 1	-0.244 (0.38)	-0.471 (0.41)	-0.263 (0.39)	-0.34 (0.43)
Belongs to CREATE		-2.491 (1.13)		-2.55 (1.18)
Current GPA			-0.321 (1.01)	-0.59 (1.15)
Confidence in teaching			-0.413 (0.86)	-0.19 (1.26)
Motivation to teach			-0.618 (1.34)	-1.41 (2.16)
Level of math anxiety			-0.442 (0.41)	-0.44 (0.42)
-2 × Log Likelihood	49.285	41.888	47.113	39.711
Change in -2 × Log Likelihood (df)		7.397 (1)	2.172 (4)	7.402 (1)
***p* value**		.007	.704	.007

Note. The subgroup sample sizes for each of the analyses were 19 CREATE cases (10 from Cohort 1 and 9 from Cohort 2) and 19 comparison cases (13 from Cohort 1 and 6 from Cohort 2).

Model comparisons are between Models 1 and 0, 2 and 0, and 3 and 2.

The estimation procedure issued warnings about model convergence similar to ones observed with logistic regressions under the descriptive results. This may be due to the very few cases not retained among Black educators in the CREATE group. The time 3 standard errors appear inflated. Results should be interpreted with caution.

Table 37 shows the "survival percentages" for the matched sample of Black educators. The results highlight the strong contrast in results between conditions for matched samples of Black educators.

TABLE 37. THE PERCENTAGES OF TEACHERS LEAVING OR REMAINING ON THEIR CAREER TRAJECTORIES AMONG MATCHED SAMPLES OF BLACK EDUCATORS

	CREATE		Comparison	
	% Left during year	% Remaining (Survival percentages)	% Left during year	% Remaining (Survival percentages)
Year 0: At start of GSU CEHD		100%		100%
Year 1: Year of GSU CEHD	2.9%	97.2%	23.5%	76.5%
Year 2: First year teaching	2.4%	94.8%	18.2%	62.6%
Year 3: Second year teaching	0.0%	94.8%	0.0%	62.6%

Chapter 8. Discussion

This quasi-experiment—funded by an i3 development grant—provides the first independent, comprehensive evaluation of the CREATE program.[18] CREATE is a teacher residency program for those aspiring to teach in local high-needs K–8 schools. The logic model posits that CREATE seeks to raise student achievement by increasing teacher effectiveness and retention of both new and veteran educators through developing critically-conscious, compassionate, and skilled educators who are committed to teaching practices that prioritize racial justice and interrupt inequities. This quasi-experiment follows two staggered cohorts of study participants (CREATE and comparison early career teachers) from their final year at GSU CEHD through their second year of teaching, starting with the first cohort in 2015–16. This study monitored the extent to which CREATE was implemented with fidelity and examined the impact of the program on several outcomes for early career teachers and their students. For teachers, we examined exploratory intermediate outcomes including measures of executive functioning, self-efficacy, commitment to teaching, and retention in teaching, as well as confirmatory outcomes including ratings of instructional strategies and positive learning environment. For students, we examined confirmatory outcomes including ELA achievement, math achievement, and general (ELA and math) achievement.

The study found that the program met fidelity for three of the five key components of the CREATE residency program—progressive core classroom roles, CF, and, SRA—for the years in which they were measured. The CBCT component did not meet fidelity for two of the three years, and the multiple forms of mentoring component was not met in either of the two years in which they were measured. Important to note is that all indicators related to the CREATE program team placing CREATE residents into their progressive classroom roles (student teacher, co-teacher in the first year of teaching, and sole teacher of record in the second year of teaching) and offering CREATE training sessions met fidelity. Components that had indicators that did not meet fidelity were those that were based on attendance of residents and mentors at training sessions. This does not mean that attendance rates were low. In fact, they were not far from meeting the high thresholds. For the CBCT component, for indicator 2 in Year 2, even though 31 out of 34 (91%) residents attended at least seven CBCT classes, the indicator did not meet fidelity because the threshold was 95% of residents. For the multiple forms of mentoring component, in Year 2, as many as 27 out of 34 (79%) residents received a full score by attending at least 12 semi-monthly meetings with their mentor for Cohort 1 and for attending at least 28 meetings for Cohort 2. However, the indicator did not meet fidelity because the threshold was 95% receiving a full score and no resident receiving a score of 0 (6 out of 34 [18%] residents received a score of 0). In Year 3, for indicator 1, 18 out of 24 (75%) of residents were paired with mentors who received prior training, while the threshold was set at 100%. Indicator 2 was just one resident short of meeting the threshold—21 out of 24 (87%) residents had a mentor who attended training during their mentor year, and the threshold was 90%.

In regard to impact, the study found that for early career teachers**,** there was not a statistically significant impact of the CREATE program across the full sample on any of the measures of executive functioning, self-efficacy, or commitment to teaching, with effect sizes ranging from -0.293 to 0.311 and p values ranging from .247 to .731. However, the impacts of CREATE on Black educators, were all positive: 0.388 (p = .175) for resilience, 0.146 (p = .460) for mindfulness, 0.370 (p =

[18] The Investing in Innovation Fund provides grants to applicants with a record of improving student achievement and attainment in order to "expand the implementation of, and investment in, innovative practices that are demonstrated to have an impact on improving student achievement or student growth, closing achievement gaps, decreasing dropout rates, increasing high school graduation rates, or increasing college enrollment and completion rates" (U.S. Department of Education, 2017).

.191) for self-efficacy in teaching, 0.026 (p = .895) for commitment to teaching, and 0.700 (p = .042) for stress management. The differential impacts on Black educators are also positive for all five outcomes, with two being statistically significant (resilience and self-efficacy) and one being marginally significant (stress management related to teaching).

In regard to teacher performance, the study found no statistically significant effect of CREATE on the TAPS (the observation component of the teacher evaluation system in Georgia) instructional strategies professional standard (ES = -0.339, p = .221) or on the positive learning environment performance standard (ES = -0.557, p = .192). However, there were specific technical issues with the TAPS measures that may render the results invalid, as we discuss below.

For students, there was no statistically significant effect of CREATE on ELA achievement (ES = -0.122, p = .454), on math achievement (ES = -0.175, p = .569), or on general (ELA and math) achievement (ES = -.139, p = .234), as measured by the Georgia Milestones Assessment System. Exploratory analyses using a larger sample based on a different matching method also yielded nonsignificant results.

Investigations into whether the impact of CREATE on teacher retention varies by subgroup reveal a promising trend that warrants great optimism for the program. Exploratory analyses showed a positive and statistically significant impact on undisrupted retention over a three-year time period (spanning graduation from GSU CEHD, entering teaching, and retention into the second year of teaching) for the CREATE group, compared to the comparison group (p = .038). We also observed that higher continuous retention among Black educators in CREATE, relative to those in the comparison group (p = .021), is a large driver of the favorable impact. The percentages of teachers in CREATE, as averaged across the two study cohorts, who maintain an uninterrupted trajectory of graduating from GSU CEHD, and taught in their first and second year are 95%, 87% and 85%, respectively. In the matched comparison group, the corresponding values are 88%, 73%, and 68%. Among Black teachers in CREATE, the values are 99%, 96% and 96%, respectively. In the matched comparison group of Black teachers, the corresponding values are 86%, 69%, and 63%.

To shed light on these findings, we offer a few working hypotheses, some of which apply across the study (such as small sample sizes), whereas others apply only to specific outcomes. We also situate our results in the literature where possible.

First, we draw limited conclusions from certain analyses of impact due to small sample sizes. During study design and recruitment, we had anticipated and factored in the estimated level of attrition into the power analysis, and we successfully recruited the targeted number of teachers. However, several unexpected limitations arose during the study that ultimately resulted in small analytic samples. These limitations included challenges in obtaining research permission from districts and schools, which would have allowed participants to remain active in the study, as well as study participants becoming ineligible to continue participation in the study due to life changes (e.g., obtaining teaching positions in other states, leaving the teaching profession completely, or feeling like they no longer had the time to complete data collection activities). Also, Georgia administers the Milestones state assessment in grades 4–8, and many participating teachers in both conditions taught in lower elementary school grades. For the analysis phase, many factors resulted in small student samples: reduced teacher samples, the technical requirement of matching of students across conditions within each cohort in order to meet WWC evidence standards, and the need to match students within grades, given the lack of vertically scaled scores.

These challenges resulted in sample sizes of only 27 (14 CREATE and 13 comparison) teachers and 29 (13 CREATE and 16 comparison) for the analysis of impact on teacher's instructional strategies and positive learning environment, respectively. For the analysis of impact on student achievement in math, the sample was as small as 52 students in each of the two conditions. We did achieve baseline equivalence between the CREATE and comparison groups for analytic

samples, but the small number of cases greatly reduces the scope and external validity of the conclusions. While we can feel confident that we are comparing outcomes for similar cases across conditions, we do so for only a small subset of the intended sample. The most robust samples were for retention outcomes, and we have the most confidence in those results.

Second, we could not detect impact on teachers' ratings on the two TAPS standards because of the lack of variability in the ratings across the sample as a whole. As mentioned in the results chapter on TAPS, the variance observed in the ordinal rating scale was remarkably low, with ratings overwhelmingly centered on the median value. The literature documents this lack of variability in teaching performance ratings. A seminal report, *The Widget Effect*, by The New Teacher Project (Weisberg et al., 2009) called attention to this "national crisis"—the inability of schools to effectively differentiate among low- and high-performing teachers. The report showed that in districts that use binary evaluation ratings, more than 99% of teachers received a satisfactory rating. In those that use a broader range of rating options, less than 1% of teachers received a rating of unsatisfactory. In effect, teachers were like widgets: undifferentiated as individual professionals.

More recent work that examined teacher performance ratings from 24 states revealed that while the full distributions of ratings vary widely across states (0.7% to 28.7% rated below Proficient and 6% to 62% rated above Proficient), the percentage of teachers rated Unsatisfactory remains less than 1% for a great majority of the states (Kraft & Gilmour, 2017). Surveys of principals in that study suggest that the main reasons for such results were "time constraints," "personal discomfort," "teachers' potential and motivation," and "challenges of removing and replacing teachers." The latter two reasons are most relevant to teachers in our study, given that the teachers are very early in their teaching career (first year teachers), and given the high turnover rate of teachers in Georgia. Principals who responded to the survey elaborated that they were more reluctant to give new teachers a rating below proficient because they acknowledge that new teachers were still working to improve their teaching, and that "giving a low rating to a potentially good teacher could be counterproductive to a teacher's development" (Kraft & Gilmour, 2017).

This implies three problems with the measure. One pertains to the validity of inferences about teacher performance based on TAPS because ratings are affected by construct irrelevant factors (e.g., concerns with implications of low or high ratings leads to score attenuation). The second involves potential for bias in estimates of impact given very low variance in scores. Specifically, the magnitudes of the standardized effect sizes are inflated by very low standard deviations. The third has to do with low power to detect differences. Intuitively, very large samples may be required to detect impacts that depend on differences in the very small proportions of ratings that are not at the median.

Third, we turn to the literature to shed light on the impact results of teacher residency programs on student achievement. A comprehensive review of teacher residency programs by The Learning Policy Institute (Guha et al., 2016) pointed out that because most residency programs are relatively new and do not yet have enough years of student achievement, few studies have been able to report on impact on student achievement. Results from a few in-depth studies of residency programs report mixed results. For example, a study of the New Visions Hunter College Urban Teacher Residency (UTR) in New York City found that in the first year of teaching, the UTR group outperformed the non-UTR group in only two of seven New York State Regents exams. An analysis of interaction effects to examine differences by years of experience and by subject showed that over time, the impact of UTR could strengthen (for geometry, algebra 2, and earth science), or diminish (for chemistry). For certain subjects, there were no significant interaction effects (Sloan et al., 2018). A report on the Memphis Teacher Residency program in 2014 found that residency graduates had higher student achievement gains than other beginning teachers and larger gains than veteran teachers on most standardized state assessments (Guha et al., 2016). Yet, another study of the Boston Teacher Residency (BTR)—a practice-based teacher preparation program in which teacher candidates work alongside a mentor teacher for a year before becoming a teacher of record in Boston Public

Schools—found that by the fourth and fifth years, BTR graduates were outperforming veteran teachers. However, it is important to note that "*initially*, BTR graduates are no more effective at raising student test scores than other novice teachers in English language arts and less effective in math" (Papay et al., 2012). These results point to the possibility that impact on student achievement might not be present or detected in earlier years of teaching. Moreover, as the Learning Policy Institute report points out, one of the limitations of these studies is the small samples, which is something we experienced in our own study, and which is problematic for drawing strong inferences as discussed above.

Last, but perhaps most importantly, this study found that CREATE teachers experienced higher retention rates in teaching over the three-year time period, compared to the comparison group, and that high retention among Black educators in CREATE is a big driver of the favorable impact. This finding is particularly important given the context of teacher retention both nationally and in the state of Georgia. At the national level, teachers are leaving the profession at an alarming rate: 44% of new teachers in public and private schools leave teaching within 5 years of entry (Ingersoll et al., 2018). Georgia experiences a similar pattern of retention, with the state's newly hired teachers leaving the workforce after their first year of teaching at an average rate of 13%, and 44% leaving teaching after five years. A 2015 study by the Georgia Department of Education, which surveyed over 53,000 teachers, reported a harsh reality:

> Teachers described a profession that was overcrowded with mandated tests, evaluated by unfair or unreliable measures, and constantly being changed without any input from the professionals inside the classroom. All occurring while being compensated poorly when time and experience are taken into account…The tens of thousands of responses displayed the effects of the current state of teaching in Georgia: a workforce that feels devalued and constantly under pressure. (Owens, 2015)

These challenges are even more pressing for Black teachers. While Black teachers once had higher retention than White teachers, Black teachers now face a very high turnover rate (22%) that is almost 50% more than that of non-Black teachers. In the South, Black teachers experience an even greater turnover rate of 26% (Carver-Thomas & Darling-Hammond, 2017). Black educators were often underprepared by teacher education programs and reported feelings of isolation, unresponsiveness from professors, and a lack of relevant coursework (Mosely, 2018). This is particularly alarming when one considers studies that show that student achievement is negatively impacted by high teacher turnover, with even stronger impacts for minority students (Ronfeldt et al., 2013).

The extant literature shows that high quality teacher preparation, alongside teacher induction programs, have demonstrated positive impacts on teacher retention. The Learning Policy Institute's comprehensive review of the key residency studies indicates that the studies consistently report that their graduates have high retention rates, ranging from rates of 80–90% in the same district after three years and 70–80% after five years. The review pointed to two rigorous studies that controlled for a range of school and district characteristics and found that there were significant differences in retention rates between residency graduates and non-residency peers (Guha et al., 2016). Similarly, the National Center for Teacher Residencies (NCTR) reported that graduates from urban teacher residency programs across the network remain in the profession, with 85% of graduates still teaching three years after graduating from NCTR's partner residencies (NCTR, 2020). These retention rates far outpaced existing research indicating that half of all teachers leave high-needs schools within three years (Allensworth et al., 2009).

The findings from our study not only corroborate the evidence of the impact of teacher residencies on teacher retention, they also shed light on the retention of Black teachers, which has not been explored widely in the literature on teacher

residencies. Our exploratory results that show statistically significant differential impacts favoring Black teachers on resilience and self-efficacy in teaching and a marginally significant differential impact on stress management related to teaching begin to chip at questions about the "why" behind these retention statistics. These findings deserve much greater attention in future research, which could include mixed-methods, longitudinal studies on teacher support and retention beyond the typical three years of most studies on teacher residencies.

As mentioned above, this study of these first two cohorts was part of an i3 development grant that began in 2015–16. Since then, the CREATE program has evolved as lessons emerged during the study and in response to the ever-changing educational and social contexts. To address existing inequities in education, CREATE has expanded equity-centered professional learning opportunities. These offerings focus on educators building critical consciousness and reflecting on the role of identity in creating equitable classrooms. In conjunction with CREATE's compassion-centered programming, these offerings are expected to develop educators who are committed to teaching practices that prioritize racial justice and interrupt inequities in student achievement.

The research on CREATE continues through this evolution, and we expect that CREATE's increasing focus on providing structure and spaces for educators to engage in racial equity work will only strengthen CREATE's impact on Black educators. As of the writing of this report, through various sources of funding, we have secured plans for studying CREATE through the eighth cohort (2022–23) of CREATE residents. In subsequent studies of CREATE, we hope to address some of the limitations discussed above, such as fidelity of implementation, small sample sizes, and teacher performance measures. We also hope to leverage the longitudinal data to explore the impact of CREATE on outcomes that may take longer to materialize, such as student achievement. Last but not least, we hope to delve deeper into the promising findings around CREATE's impact on teacher retention, particularly retention of Black teachers, and explore the factors that may moderate and mediate these impacts. Given the encouraging findings that are emerging from studies of teacher residency programs in general, as well as from this very study, we hope that further research will continue to inform CREATE as it strives toward the vision of improving student academic and social emotional growth through critically-conscious, compassionate, and skilled educators who are committed to teaching practices that prioritize racial justice and interrupt inequities.

References

Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The Schools Teachers Leave: Teacher Mobility in Chicago Public Schools*. Consortium on Chicago School Research.

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. Assessment, 13, 27-45

Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005). Using Experiments to Assess Nonexperimental Comparison-Group Methods for Measuring Program Effects. In H. S. Bloom (Ed.), *Learning more from social experiments*, 173-235. Sage Foundation

Carver-Thomas, D. & Darling-Hammond, L. (2017). *Teacher turnover: Why It Matters and What We Can Do About It*. Learning Policy Institute.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Davidson, J. R. T., & Connor, K. M. (2016). Connor-Davidson Resilience Scale (CD-RISC) Manual. Unpublished. http://cd-risc.com/

edTPA for Georgia. (2020). Pearson Education, Inc. Retrieved from https://www.edtpa.com/PageView.aspx?f=GEN_Georgia.html

Elder, T., Wang, J., & Cramer, S. E. (n.d.). *An Evaluation of the Validity and Reliability of the Intern Keys Assessment*. University of Georgia College of Education. https://gaedassessment.files.wordpress.com/2018/05/intern-keys-validation-project-final-report-9-23-15.pdf

Elfers, A. M., Plecki, M. L., & Knapp, M. S. (2006). Teacher Mobility: Looking More Closely at "The Movers" Within a State System. *Peabody Journal of Education 81*(3):94-127. https://www.researchgate.net/publication/240519657_Teacher_Mobility_Looking_More_Closely_at_The_Movers_Within_a_State_System

Georgia Professional Standards Commission (2019). Retrieved from https://www.gapsc.com/Certification/TieredCertification/induction.aspx

Georgia Professional Standards Commission (2014). *Check Certification Status*. https://www.gapsc.com/Certification/Lookup.aspx

Georgia Department of Education (GaDOE). (2020). Teacher Keys Effectiveness System. https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Pages/Teacher-Keys-Effectiveness-System.aspx

Georgia Department of Education (GaDOE). (2019). *An Assessment & Accountability Brief: 2018-2019 Georgia Milestones Validity and Reliability*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Milestones/2018-19_Georgia_Milestones_Validity_and_Reliability_Brief.pdf

Georgia Professional Standards Commission (GaPSC). (n.d.). Induction. https://www.gapsc.com/Certification/TieredCertification/induction.aspx

Guha, R., Hyler, M. E., & Darling-Hammond, L. (2016). *The Teacher Residency: An Innovative Model for Preparing Teachers*. Learning Policy Institute.

Ho, D. (2005). *Matchit: Matching Software for Causal Inference* [Computer software]. https://www.rdocumentation.org/packages/MatchIt/versions/1.0-1/topics/matchit

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis 15*(3), 199-236. http://gking.harvard.edu/files/abs/matchp-abs.shtml

Ingersoll, R. M., Merrill, E., Stuckey, D., & Collins, G. (2018). Seven Trends: The Transformation of the Teaching Force – Updated October 2018. CPRE Research Reports. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1109&context=cpre_researchreports

Kraft, M.A. & Gilmour, A.F. (2017). Revisiting the Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234-249

Mosely, M. (2018). The Black teacher project: How racial affinity professional development sustains Black teachers. *The Urban Review*, *50*(2), 267–283. https://doi.org/10.1007/s11256-018-0450-4

National Center for Teacher Residencies (NCTR). (2018). *2017 Stakeholder Perception Report*. https://nctresidencies.org/wp-content/uploads/2018/06/June-2018_NCTR-Stakeholder-Report-Final.pdf

Open Georgia. (2008). State of Georgia Financial Reports. http://open.georgia.gov/

Owens, S. J. (2015). *Georgia's Teacher Dropout Crisis: A Look at Why Nearly Half of Georgia Public School Teachers are Leaving the Profession*. Georgia Department of Education. https://www.gadoe.org/External-Affairs-and-Policy/communications/Documents/Teacher%20Survey%20Results.pdf?utm_source=rss&utm_medium=rss

Papay, J. P., West, M. R., Fullerton, J. B., & Kane, T. J. (2012). Does an urban teacher residency increase student achievement? Early evidence from Boston. *Educational Evaluation and Policy Analysis*, *34*(4), 413-434.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American educational research journal*, *50*(1), 4-36.

Singer, J. D., & Willett, J. B. (1993). It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics, 18*(2), 155-195.

Sloan, K., Allen, A., Blazevski, J., Carson, F., & Rockman, S. (2018). *A Different, More Durable Model: Hunter College Urban Teacher Residency Project*. New Visions for Public Schools.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. The New Teacher Project. http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

U.S. Department of Education. (2017). *Investing in Innovation Fun*d. https://www2.ed.gov/programs/innovation/index.html