NORTHWESTERN UNIVERSITY

Child and Family Policy in the 21st Century: A Focus on Early Childhood Education and Parental Work

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Human Development and Social Policy

By

Olivia Healy

EVANSTON, ILLINOIS

September 2021

Abstract

This dissertation consists of three studies related to parental employment, maternity leave policy, and early care and education.

In study 1, my coauthor, Jennifer Heissel, and I evaluate the impact of parenthood on men and women's job performance and career advancement using detailed data from the U.S. Marines. For parents who remain employed after having a baby, disruptions in home life and health may spill over into their performance at work. Using monthly data from 2010 to 2019, we exploit variation in the precise timing of first births to identify impacts on health-dependent measures of worker performance. We then compare parents' promotion trajectories to similar non-parents' trajectories, using a matching approach that assigns non-parents to "placebo births." We find negative impacts on parents' employer-assessed physical fitness and supervisor-rated job performance, concentrated mainly among women. Consistent with these findings, women's promotion trajectories slow down in response to childbirth but men's do not. In a complementary analysis, we exploit sudden policy changes to the length of paid maternity leave to explore whether leave length is associated with mothers' job performance and career advancement. Longer leaves exacerbate declines in women's job-related physical fitness but do not consistently appear to slow their promotion trajectories to a greater or lesser degree. Results suggest longer periods away from work due to maternity leave may erode job-specific skills but not on the margin that drives changes in career advancement in this setting.

In study 2, along with coauthors Kathryn Gonzalez, Luke Miratrix and Terri Sabol, I explore variation in observer-rated preschool teaching quality within the school year and implications for the accuracy of early childhood education accountability evaluations. Accountability systems designed to monitor and enhance early childhood education program quality increasingly rely on observational ratings of teachers' skills. We draw on a unique data set with 2,803 observational ratings of 303 preschool teachers to characterize within-school-year patterns of growth and decline in teacher quality, as measured by the observational tool the Classroom Assessment Scoring System (CLASS^{TM}) (Pianta et al., 2008). We then quantify the share of total within-teacher variability in ratings explained by time trends. Last, we simulate accountability outcomes under the federal Head Start accountability policy, based on the time of year a program's teachers are assessed. Results show observed teacher quality ratings generally decline during the beginning of the school year; improve during the winter months; and plateau in the spring. This pattern is particularly pronounced for instructional support CLASS ratings, which capture how teachers foster students' higher-order thinking. Simulation analyses confirm sizeable differences in accountability outcomes based on program observation date, with especially large differences in the likelihood of failing a Head Start accountability review due to time-patterned variation in instructional support scores.

In study 3, I investigate how early care and education providers respond to the expansion of public pre-Kindergarten (pre-K) funding in Illinois. Federal, state, and local governments have increasingly invested in such programs, with the goal of increasing families' access to early care and education settings. However, it remains unclear whether public pre-K funding effectively draws new providers into the market for care and, if not, whether already established providers that receive public pre-K funds shift program operations in response. Using detailed longitudinal data on the universe of child care providers in Illinois, I identify

what portion of public pre-K providers that receive funding are new to service delivery vs. already established. Next, among established providers, I compare changes in service delivery before and after public pre-K funding receipt to changes over time in observably similar providers, identified using a propensity score matching approach. I find Illinois' pre-K expansion funds go to very few new providers; 75% of funded programs existed for 3 years or more before they were awarded funding. Among these, public pre-K funding increases the odds that a provider remains in business and increases the number of distinct child care sessions it offers for preschoolers. The latter finding is closely related to the nature of the funding, which covers only part-day care for eligible children. Findings on the immediate impacts of public pre-K on funded providers begin to inform our understanding of how the early care and education sector responds to state pre-K expansion.

Acknowledgments

I am thankful to my family, friends, and colleagues who helped me set me on the path to graduate school and made it possible for me to complete this dissertation. First, I would like to thank my dissertation committee: Jon Guryan, Terri Sabol, Diane Schanzenbach, and Matt Notowidigdo. Your feedback challenged me throughout the process and improved my studies greatly. I am grateful for countless advising meetings, professional advice, and personal sup¬port as I learned to become an independent scholar. I also thank David Figlio who formally advised me early in graduate school and provided continued mentorship thereafter. Thanks too go to Kirabo Jackson, Ofer Malamud, and Hannes Schwandt for feedback in HDSP Econ Lab and at various other junctures.

Early in graduate school, I learned the value of connecting with and learning from my peers. In this regard, I am grateful to Jenni Heissel—who has become a mentor, collaborator, and friend. I am indebted to Jenni for her professional wisdom, excellent pep talks, endless supply of research-related Tweets, and amazing children whose company makes for the best work breaks. Next, I must thank Ryan Svoboda, a great officemate and friend who kept me laughing with spontaneous Ryan-rants and gave me countless rides home in his trusty Subaru. I am grateful too for other senior HDSP colleagues for showing me the way forward. This includes Emily Ross, Mollie McQuillan, Andrea Busby, and Courtenay Kessler. Last, CC DuBois's professional advice got me through difficult moments in graduate school. CC also taught me that dedicating time for friends (e.g., watching a movie together on Sundays) could be more important than doing work, a lesson that has paid dividends. I miss you, CC.

My current peers in HDSP and at Northwestern have also been major pillars of sup¬port. First, I thank Cora Wigger, my partner in all things econ-oriented and job market-related. Regular Monday and Friday zoom meetings with Cora during the global pandemic were

instrumental in helping me make it to the finish line. I am also grateful to Heather McCambly, Sheridan Fuller, Claire Mackevicius, Timi Viragh, Jen Cowhy, Naomi Blauschild, Ayah Kamel, and Sarah Peko-Spicer. They helped improve my research, supported me with "tomato timer" work sessions over zoom, and included me in meaningful dialogue and action around inequities in our community, university, and society at large.

Anika Nerella provided excellent research assistance by geocoding and cleaning the data in this dissertation on child care and public Pre-K providers, among other tasks. Joellyn Whitehead made it possible for me to learn from these data by carving out time in her busy schedule at the Illinois Network of Child Care Resource and Referral Agencies to query the archives for needed records and support my work. I am also thankful for staff at the Illinois State Board of Education and Illinois Early Childhood Asset Map who provided me with data. I am indebted to current and former students at the Naval Postgraduate School for their pro ject assistance and patient responses to every last question I have ever had about military life. This includes Capt. Mike Larson, Ma j. Tamara Cordero, Capt. Julie Schumacher, and Capt. Amanda Henegar. It has been a pleasure to work with each of them.

I have also benefited from generous financial support over the course of my graduate career. I am grateful for the funding provided by the Office of Planning Research and Evaluation, an Office of the Administration for Children and Families within the Department of Health and Human Services; the Institute of Education Sciences' Multidisciplinary Pro¬gram in Education Sciences; The Graduate School at Northwestern University through their Graduate Research Grant; and the School of Education and Social Policy at Northwestern University through the Robert Cowell Fellowship.

My friends outside of my immediate academic sphere have played pivotal supporting roles. Rebecca Gourevitch is a constant open ear; we discuss the most meaningful personal

events in our lives as well as the details of dissertation progress in turn. The combination of a dear friend turned colleague in Rebecca has been a true gift. Rachel DeLevie-Orey has offered deep, reassuring empathy like no other person I know, plus the needed levity and wit to get me through. Leah Abrams has been a steady support in my life for years, seeing me from high school through college, my first job, graduate school and now off to a postdoc. Chrissy Boyd spent hours with me in study lounges and the library, making my time at Cornell an absolute joy. Finally, my dearest childhood friend, Allison Portenoy, helped me survive college and make it to/through a degree in policy analysis and management. I would not have gotten here without her. Thank you, Port.

My Aunt Judy and Uncle Francesco made college, among many other opportunities, possible for me. I am deeply grateful to them. I must also thank Susan Sussman who helped me apply to college and introduced me to the field of public policy by suggesting it as a major. The suggestion clearly stuck! Thanks too to my undergraduate advisor, Rachel Dunifon, who helped launch my research career, validating the idea that I could go onto complete a PhD.

My grad school cohortmate and now good friend, Emily Hittner, also played a huge role in helping my complete this dissertation. She was a cheerleader for me when I needed it and helped me find joy and laughter during hard times. She was also adept at pointing at my computer screen when I was online shopping instead of working. Emily has offered me advice on virtually every dilemma I have faced over the past six years, and she has become one of my most trusted sources of support. Thank you, friend.

My father has also helped propel me to this final educational juncture. I know no one smarter, harder working, nor more willing to lead by example than my father. He earned his college degree the year I graduated from high school. Then, he went on to earn his law

degree the week before I got my bachelor's. Thank you for teaching me to strive, Dad.

My stepmom Karen has offered invaluable insight into and empathy for this PhD process. She is my constant source of support and helps make sure I stay afloat. It has meant the world to know Karen is looking out for me, even from miles away, ready to shoulder some of my burden whenever she can.

Finally, my husband Scott Monsky has been my greatest champion. He moved from the beaches of Florida to the subzero temperatures of Chicago so that I could pursue this degree. Scott has celebrated every academic milestone and commiserated any setbacks with me too. There is no challenge he has not been willing to help me see through. Scott, I am endlessly grateful to have you as a partner. You take such good care of "our family" and have sacrificed so much to support me in my career. Thank you, and I love you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the past few decades, the United States experienced a major shift in maternal employment rates. The proportion of women with young children in the labor market rose from 34% of mothers with children under age three in 1975 to 64% percent in 2019 (Bureau of Labor Statistics, 2020). Now, in most families – whether headed by two parents or a single parent – all parents work (Bureau of Labor Statistics, 2020). Despite this shift in family life, the design, implementation, and study of U.S. policies to support working parents is only beginning to catch up. In contrast to other developed countries, the U.S. lacks a universal paid leave policy after childbirth (Organization for Economic and Co-operation Development Family Database, 2019). Parents in the U.S. report access to reliable, affordable child care is hard to come by and, regardless of financial background, parents find it difficult to balance the demands of work with the responsibilities of family (Pew Research Center, 2015)

Without access to key supports, such as child care or the ability to take leave from work, parents may struggle to remain in the labor market. Women may be particularly affected, given that for decades women were expected to manage family responsibilities and remain out of the labor market. This dissertation brings together research at the intersection of

parenthood, policy, and employment to consider the role that children play in parents' labor market experiences and the ways public policy supports for parents and children might more effectively respond.

Study 1 focuses on the impact of parenthood itself on mothers' and father's job performance and career advancement. The coauthored paper leverages novel data from the U.S. Department of Defense (DoD) to assess changes in parents' job-related physical performance, supervisor job proficiency ratings, and promotion trajectories among first-time, active-duty Marine parents. We then use exogenous shocks to the length of DoD-funded parental leave to examine whether more generous leave policies mitigate (or exacerbate) the effects of parenthood on work performance. Previous literature has typically examined child-driven gender gaps in job productivity and career advancement among more educated professionals with easy-to-observe outcomes, such as advancement to senior management or academic tenure (Antecol et al., 2018, Bertrand et al., 2010, Kim and Moser, 2020). Our study provides evidence on a more diverse group of workers not currently represented in the literature – Marine Corps service members. Marines come from a variety of financial and racial/ethnic backgrounds and work across a range of occupations, many common to the civilian workforce. Our study leverages monthly data and is the first, to our knowledge, to trace month-by-month dynamic work performance and promotion responses to parenthood. The study also sheds light on the role of maternity leave length in shaping parents' job performance outcomes.

The second and third studies turn the focus to early care and education policy – how to improve its quality and expand families' access to care. In study 2, my coauthors and I investigate whether preschool classroom quality evolves over the school year and how predictable patterns of growth and decline complicate government efforts to monitor and improve early

education quality. We leverage rich data on preschool teachers and their observed quality ratings, assessed at multiple time points throughout the preschool year. Our general analytic approach relies on multilevel linear piecewise regressions to model within-year time trends in quality ratings. We then decompose the share of total within-teacher variability in ratings explained by time trends to quantify their magnitude. Based on our findings, we simulate how likely early childhood programs are to be misclassified under accountability systems given time-patterned trends in quality ratings and different program-specific observation dates. With early childhood education accountability policies increasingly relying on observational measures of classroom quality, results from study 2 help inform the design of accountability policies that incorporate these measures.

Study 3 explores how expansions in public pre-Kindergarten (pre-K) programs interact with the already-established infrastructure for early care and education. Specifically, I study how increases in funding for public pre-k for low-income and otherwise socioeconomically disadvantaged families in Illinois implicate early care and education service provision among funded providers. First, I explore whether the expanded public pre-K services take place in new or existing early education facilities. Second, among established providers, I investigate whether public pre-K funding impacts program operations, including the propensity to stay in business, expand capacity, or shift other types of services offered to families. I employ a propensity score matching approach to approximate counterfactual outcomes for public-pre-K-funded providers. I find that public pre-K funding increases the likelihood that a child care provider remains in business, but it does not appear to increase their capacity to serve more children. If anything, early education facilities that receive public pre-K funding increase the number of sessions they offer within a single day, suggesting they may serve more children on net but with shorter periods of care.

This dissertation research is especially timely. The COVID-19 pandemic has underscored tensions parents, particularly mothers, face in balancing family responsibilities and work in the face of inadequate access to child care and other work supports due to the pandemic. Moreover, each of these studies offers relevant evidence to inform recent policy proposals, such President Biden's American Families Plan, which call for the creation of a comprehensive paid family leave program for parents, increased access to high-quality affordable child care, and universal public pre-K for children starting at age 3.

Chapter 2

Baby Bumps in the Road: The Impact of Parenthood on Job Performance and Career Advancement

2.1 Introduction

Almost four million babies are born in the U.S. every year (Hamilton et al., 2019). While having a child can be an exciting milestone, it affects a multitude of social, economic, and behavioral health outcomes for parents. In terms of health, parents with newborns experience sleep deprivation, increased emotional stress, and changes in their neurobiology (Saxbe et al., 2018). For women in particular, pregnancy and childbirth are major medical events. Women experience immediate risk of infection, complications, and postpartum depression after having a baby (Memon and Handa, 2013, O'Hara and Swain, 1996), and physiological recovery takes on average 1 year (Melchiorre et al., 2016).

Despite the ubiquity of parenthood, we have limited evidence on how the mental and physical strain of having a baby implicates parents at work, particularly in terms of their on-the-job performance. Prior research has necessarily focused on new parents' labor force attachment, hours worked, or wages. These are often the only measures available, and mounting evidence shows women increasingly leave the labor market, cut back on hours worked, and earn lower wages after having a child (Agüero and Marks, 2011, Angrist and Evans, 1998, Bronars and Grogger, 1994, Cáceres-Delpiano, 2006, Cools et al., 2017, Cruces and Galiani, 2007, Jacobsen et al., 1999). These changes drive persistent earnings penalties that accrue to mothers, but not fathers, over decades (Angelov et al., 2016, Barth et al., 2017, Bertrand et al., 2010, Kleven et al., 2019a,b). One possible explanation for these findings is that mothers either prefer to spend more time on caregiving or face more societal pressure to provide care than fathers. Another possible explanation is that job demands become overly taxing after having a baby, perhaps especially so for women such that they opt to stay home or move into lower paid, more flexible positions in response to motherhood.

In the current paper, we explore this second possibility, considering whether new parents become less physically able to perform on the job. We then investigate whether they experience slower career advancement after having a child and if access to longer maternity leave helps offset any negative consequences of parenthood for mothers. The primary advantage we have in addressing these research questions is access to detailed, consistently measured, longitudinal data for individuals. Data come from Department of Defense (DoD) administrative records on service members in the U.S. Marine Corps, recorded monthly covering 2010 to 2019. They include high-frequency, repeated measures of worker performance, including job-related physical fitness tests and supervisor evaluations of job proficiency.

We rely on two empirical strategies to isolate the causal effect of childbirth on parents' job

performance and career advancement, exploring impacts separately for men and women. One approach uses an event study framework and leverages the precise timing of childbirth as an exogenous shock to parents' work performance outcomes. Under this event study approach, we use data on same-gender Marines who do not have a child during the study window to account for secular time trends in our outcomes. We observe minimal evidence of pretrends in outcomes before birth, lending confidence to the research design. Our second approach compares first-time parents' promotion trajectories to similar nonparents' trajectories, using a matching strategy that assigns nonparents to "placebo births." We then examine whether gaps in career advancement emerge after a birth vs. a placebo birth event. We explore variation in impacts on women's job-related physical fitness and promotion trajectories across DoD maternity leave policies that vary in length. Two major policy changes during our study window create plausibly exogenous variation in the length of paid leave available to women.

While the military differs in obvious ways from many professions, aspects of the DoD context make it especially suitable for this analysis. First, we capture a diverse group of workers. Individuals who enlist in the Marines work in a range of occupations, including ones common to the civilian workforce, such as food service, traffic management, information technology, and more. Second, effective job performance in the Marines is tightly linked to physical health. Regardless of their occupational specialty, Marines need to be physically ready to support combat missions at any time. Although this professional requirement may seem unique, nearly half (45%) of all jobs in the civilian labor market require at least medium physical strength, defined as work that involves frequent lifting or carrying of objects weighing up to 25 pounds (Bureau of Labor Statistics, 2017a). Moreover, the average civilian worker stands or walks for 60% of the day (Bureau of Labor Statistics, 2017b), requiring a baseline level of physical health. As such, the Marine context offers an opportunity to

learn how parenthood impacts work performance for those in jobs that require both mental and physical acuity. Third, Marine Corps job performance measures are standardized across all occupation types. Whether a Marine works in food service or as a lawyer, he or she is regularly rated on the same performance metrics to determine promotion. Fourth, Marines generally commit to 3- or 4-year contracts, which limits selection out of work after having a child, especially compared to other employment contexts.

A final benefit of the DoD context is the opportunity to study the kind of universally accessible, fully paid parental leave not widely available publicly or privately in the U.S. The small number of public, state-level paid leave programs that exist in the U.S. generally cap wage replacement (Rossin-Slater et al., 2013). Among private firms, about half of leave is fully paid, a third is partially paid, and the remainder is unpaid (Donovan, 2019). In cases where leave is not fully paid, the decision to return to work may be driven by financial need, particularly for lower-income women (Rossin-Slater et al., 2013). In our study context, women receive full pay, and this generally means take-up rates are high. Also in our context, the length of paid maternity leave ranges from 6 to 18 weeks, well within the scope of expansions considered by policymakers at the federal- and state-levels in the U.S.

Using the event study approach, we find small but meaningful impacts of the transition to parenthood on job performance, concentrated mainly among women. As is perhaps expected, physical performance among mothers declines to the greatest degree closest to the birth. However, mothers' physical performance remains below prepregnancy levels even 2 years after the birth. In terms of supervisor ratings, mothers receive progressively lower ratings each month that passes after childbirth. Declines are small in magnitude, dropping 0.02 standard deviations per month relative to performance before pregnancy.

In contrast, we observe minimal impacts of parenthood on fathers' job performance. The

birth of a child leads to short-lived declines in men's physical performance. Fathers score 0.07 standard deviations below their prepregnancy levels on fitness tests 1 month after the baby arrives. By the child's 1st birthday, men's physical fitness performance is rated approximately the same as it was before the mother's pregnancy. Findings show fathers' supervisor-rated job performance is not meaningfully affected by the transition to parenthood.

Consistent with the persistent negative impacts of parenthood on women's job performance, we find evidence that women's promotion trajectories slow as a result of having a child. Men's promotion trajectories are unaffected by the transition to parenthood.

Across maternity leave policies, we find longer paid leave exacerbates performance declines in physical fitness but does not consistently predict smaller or larger delays in career advancement for women after having a child.[1] Women who receive 6 weeks of maternity leave recover to their prepregnancy levels of job-related physical fitness within 2 years of childbirth. In contrast, women who receive more than 6 weeks of maternity leave perform below their prepregnancy fitness levels even 2 years after childbirth, on the order of magnitude of 0.15 to 0.39 standard deviations lower (depending on the length of extended leave).[2] Despite recovering in terms of physical performance by 2 years after giving birth, mothers with the shortest leave show the most consistent promotion delays during the same time period. Results suggest that longer periods away from work may exacerbate declines in

[1]We cannot estimate the impact of leave length only on supervisor-rated job proficiency evaluations. Data on supervisor ratings are only available for Marines in service as of October 2017, at which time leave changes have already gone into effect.

[2]While longer maternity leave exacerbates physical performance declines for women in our sample, this does not rule out that longer leave might have positive effects in other domains, such as mothers' health or children's well-being. For example, Balser et al. (2020) find similar maternity leave extensions for women in the Army and Air Force reduce postpartum depression diagnoses among mothers. In other U.S. and international contexts, evidence shows access to paid parental leave improves mothers' physical and mental health (Butikofer et al., 2018, Bullinger, 2019), duration of breastfeeding (Pac et al., 2019), infant health (Bullinger, 2019), and time spent with children upon return to work (Trajkovski et al., 2019, Bailey et al., 2019).

job-specific physical performance, but physical performance changes do not appear to drive delays in career advancement for women in response to motherhood.

Our study contributes to a longstanding literature on the impact of fertility on parents' employment outcomes in a few key ways. First, we directly measure actual performance on the job—a key precursor to worker output and productivity, as well as to any changes in employment, hours worked, and wages. Second, while research has largely focused on female labor supply and highlighted motherhood as a turning point in women's careers, we explore employment impacts on both women *and* men who work in the same setting and remain employed after having a child. Many papers consider fathers as a comparison for mothers when estimating child wage penalties (e.g., Angelov et al. 2016, Bertrand et al. 2010, Kleven et al. 2019a) and family policy impacts on mothers (e.g., Balser et al. 2020). If fathers are uniquely impacted by the transition to parenthood, such estimates may over- or underestimate the effect of parenthood or policy, depending on how fathers are affected.

Third, we examine employment consequences of childbirth month by month and trace the dynamic responses of men and women *within* the 1st and 2nd years of work after becoming parents. Prior papers that estimate child penalties use annual earnings or income and therefore cannot detect immediate, within-year impacts of childbirth (Angelov et al., 2016, Barth et al., 2017, Bertrand et al., 2010, Kleven et al., 2019a,b). Our finding that parenthood impacts performance *within* the 1st year after birth points to one mechanism through which child penalties may arise. Persistent declines in mothers' health-related ability to perform at work may lead women–more so than men–to exit the labor market, cut back hours worked, or receive lower wages by the time these outcomes have been measured 1 year postbirth in other papers. Last, our paper adds to a growing literature on the impact of paid maternity leave on women's labor market outcomes. We contribute by exploring variation in length

of paid leave on new dimensions of employment, on-the-job work performance and career advancement.

The rest of the paper proceeds as follows. Section 2 provides institutional background on the DoD and details on employment, parenthood, and paid family leave in the Marines. Section 3 provides details on our data. Section 4 describes our empirical strategy, including estimation and identifying assumptions. We describe our results in Section 5 and conclude in Section 6.

2.2 Institutional Background

The DoD is the world's largest employer, with a total of 1.3 million active-duty service members across four branches – the Army, Navy, Air Force, and Marine Corps. Each branch plays a unique role in maintaining U.S. security and peace. In this paper, we focus on service members in the Marines Corps, where administrative records on job performance are readily available. The Marine Corps is an immediate response force, ready to deploy quickly to support combat missions on sea or land. Marine Corps service members make up nearly 15% of active-duty forces and have roughly 185,000 active-duty service members (Department of Defense, 2018).

2.2.1 Jobs and Associated Work in the U.S. Marines

Individuals begin in the Marine Corps either as a junior enlisted, akin to an entry-level civilian worker; or as an officer, akin to a civilian manager. Enlisted service members must have a high school degree and be between 18 and 29 years old when they begin. Officers must have at least a bachelor's degree upon entry. Roughly 89% of Marine Corps service members are

enlisted, and given age requirements, most are of prime childbearing age. Among Marines, 81% are under 30 years old, and more than one quarter are parents (Department of Defense, 2018).

In total, there are over 35 career fields in the Marines, and each has dozens of specializations, referred to as Military Occupational Specialties. Some career fields specific are specific to the military. These include infantry, field artillery, and terminal attack control (in charge of communicating with aircraft to guide offensive air operations). Other career fields are also present in the civilian labor market. These include fields such as food services, financial management, military police and corrections, legal services, and even music (e.g., the Marine Corps band). An individual can enlist in the Marines under a career field, but he or she is not guaranteed a specific job specialization within that field.

A service member's occupational specialty along with their assigned unit determines their day-to-day work environment. Our analysis focuses on active-duty Marine parents who work full time, Monday through Friday. For these individuals, the workday typically begins with early morning physical training (as early as 5:30 a.m.), followed by work assignments through the evening. Most service members live and work on or near a military base and are stationed in the U.S. (83% of service members; Department of Defense 2018), though some are stationed abroad.

Marines sign a legally binding contract that outlines their required length of service. For enlisted service members, initial contracts typically require 4 years of active-duty service, while for officers the commitment is typically 3 years.[3] Importantly for our purposes, these contracts limit the extent to which Marines can exit the labor force after they have a baby.

Effective job performance in the Marines requires both mental and physical acuity. The

[3]Contracts can stipulate additional service beyond these minimums, requiring additional years of service in the Marine Corps Reserves, which allows service members to work part time.

Marine Corps uses a standardized set of measures to evaluate performance among both active-duty and reserve Marines. Performance measures include supervisor-scored job proficiency evaluations, physical ability assessments, and a rifle marksmanship test.

Supervisors evaluate Marines for performance purposes using one of two scales, depending on the Marine's rank. Junior enlisted receive proficiency and conduct marks ("ProCons") at least twice per year, and senior enlisted and officers receive Fitness Report ("FitRep") scores at least annually. Both assessments require supervisors to rate a Marine's performance across a range of professional domains.[4] As such, these evaluations capture job proficiency. To rate physical ability to perform on the job, Marines take two tests per year: the physical fitness test (timed running, crunches, and upper-body strength) in the 1st half of the year and a combat fitness test (timed running, a combat-related obstacle course, and upper-body strength) in the 2nd half of the year. Last, rifle marksmanship assessments occur annually to evaluate Marines on their shooting skills.

Scores on each of these measures give rise to a composite performance score for certain Marines, and composite scores are updated every quarter to incorporate new performance assessments.[5] Of these three measures, supervisor performance ratings influence promotion outcomes most heavily. Marines can clearly determine what they need to advance and, therefore, have especially strong incentives to perform well on measured performance assessments.

[4]Domains include mission accomplishment (job-specific aptitude, competence, technical knowledge, and practical skills), character (courage, effectiveness under stress, and initiative), leadership (setting the example, communication skills), and intellect and wisdom (professional military education, decision-making ability, and judgement), among others.

[5]The composite score determines promotion for all ranks below E5, conditional on meeting requirements for minimum time in service and minimum time in the current job level to be eligible for promotion. Promotions at lower ranks, E1 through E3, are relatively automatic after a given number of months in service and months in rank. For promotion at ranks above E5, the same performance metrics, along with supervisor ratings, are reviewed by an evaluation board to determine promotion. As such, Marine Corps promotions are similar to civilian promotions. Both are based on work performance, but the Marine Corps promotion system is arguably more exact.

2.2.2 Parenthood and Paid Leave in the U.S. Marines

The DoD provides military parents with a number of family-friendly benefits, including fully paid parental leave. We focus on policy changes to paid leave for primary caregivers (most often women) and refer to this leave as maternity leave.[6]

Prior to 2015, all DoD branches provided active-duty women with 6 weeks of paid maternity leave. In July 2015, the Secretary of the Navy announced that primary caregivers (most commonly women) in the Navy and Marine Corps would be entitled to 18 weeks of leave. Women who had given birth earlier in the year (as of January 2015) could retroactively take advantage of the 18-week leave policy. For women who had already taken 6 weeks of leave and returned to work, they tended to use the additional 12 weeks of paid time off discontinuously. For women who were on leave or pregnant at the time of the announcement of expanded leave, the majority took the additional leave consecutively, as did mothers who became pregnant after the announcement (Bacolod et al., 2020). We analyze these two groups separately. In early 2016, the Secretary of Defense standardized maternity leave to 12 weeks for all services.[7]

[6]Changes to paternity leave (i.e., leave for "secondary caregivers") also occurred during this time. However, paternity leave is limited in scale; expansion of paternity leave in the Marines changed from 10 to 14 days. We do not focus on paternity leave impacts in this paper, given the small magnitude of the change.

[7]For Marine Corps mothers, this meant their paid maternity leave was reduced from 18 to 12 weeks. However, the policy change for Marines only applied to pregnancies that began 31 days after the announcement (i.e., to pregnancies that began on March 3, 2016 or later, per doctor estimation). As a result, Marine women who became pregnant after the policy announcement were aware they would receive 12 rather than 18 weeks of paid leave. Given that only 1 month's notice was given for the policy change, many of the women who became pregnant at the beginning of the 12-week policy change likely made the decision to become pregnant before the policy announcement.

2.3 Related Literature

A large body of literature explores the impact of fertility and, more recently, paid maternity leave and other family support policies on mothers' labor market outcomes. Prior research on children and maternal employment commonly exploits variation in family size due to twin births or third births that result from having two prior children of the same sex. These studies consistently find that an additional child reduces employment and hours worked among mothers, both in the U.S. and other countries (Agüero and Marks, 2011, Angrist and Evans, 1998, Bronars and Grogger, 1994, Cáceres-Delpiano, 2006, Cools et al., 2017, Cruces and Galiani, 2007, Jacobsen et al., 1999). Studies tend to find that fathers' labor force attachment is either unaffected by additional children (Angrist and Evans, 1998, Cools et al., 2017) or fathers' outcomes are not explored (Agüero and Marks, 2011, Bronars and Grogger, 1994, Cáceres-Delpiano, 2006, Cruces and Galiani, 2007, Jacobsen et al., 1999).

A related but more recent strand of literature investigates how career interruptions associated with motherhood affect women's total labor market earnings across the life cycle. Mounting evidence reveals a "child penalty" in earnings accrues to mothers but not fathers in response to childbearing (Angelov et al., 2016, Barth et al., 2017, Bertrand et al., 2010, Kleven et al., 2019a,b). The earnings penalty stems from reductions in labor force participation, hours worked, and wages after women have their first child (Kleven et al., 2019b).

Most similar in spirit to our paper is a set of studies focused on changes in mothers' and fathers' job productivity across the transition to parenthood. For example, Azmat and Ferrer (2017) find that female lawyers with young children are less productive compared to male lawyers with young children in terms of hours billed annually, a key productivity measure in the legal profession. Kim and Moser (2020) similarly show that female scientists in the 1950s patented less during their childbearing years. They posit lower patenting productivity

drives lower rates and slower speed of promotion to tenure for female scientists with children as compared to fathers and other women without children. Last, Gallen (2018) explores how firm output in Denmark varies by the gender and parenthood status of employees in private firms. She finds mothers are substantially less productive–measured according to a firm production function model–than other workers (nonmothers, fathers, and nonfathers), particularly during their childbearing years. By contrast, our paper homes in on a precursor to work output: work performance. Our results also provide evidence on a more diverse group of workers not currently represented in the literature.

Limited research explores the potential for family support policies to mitigate the health and work consequences of the transition to parenthood, given the difficulty of tracking men and women's outcomes over time and across different policy settings. To date, most studies focus narrowly on how access to paid family leave impacts women's labor market attachment with mixed findings. Early studies in select U.S. states show the introduction of paid leave improves women's labor force attachment and wages following birth (Baum and Ruhm, 2016, Byker, 2016, Rossin-Slater et al., 2013). However, more recent research reveals negative effects of paid maternity leave on women's employment and wages in the long term (Bailey et al., 2019, Timpe, 2019). Among first-time mothers specifically, Bailey et al. (2019) show that an additional 6 weeks of paid leave reduces a new mother's likelihood of returning to work as well as reduces her annual wages in *both* the short and long run. The authors' innovative research design exploits month of birth during the rollout of California's paid leave policy to compare women who were more or less likely to have access to leave directly after having a baby.

Beyond studies focused on maternal labor market attachment and wages, Andresen and Nix (2020) and Kleven et al. (2020) explore whether paid family leave helps narrow child

penalties in earnings. Both studies find minimal evidence to that effect. Using data on academic economists from 1980 to 2005, Antecol et al. (2018) show that family-oriented university policies can exacerbate gender-gaps in publication output and time to promotion. In their setting, gender-neutral tenure clock stopping policies designed to accommodate lower productivity among academics in the year after childbirth increase men's likelihood of receiving tenure and decrease women's. In terms of mothers' health, few studies focus on how paid leave availability shapes adult health outcomes (for an overview, see Rossin-Slater and Uniat (2019)).

To begin to fill the gaps in the current literature, the present study explores the impacts of the transition to parenthood on men and women's physical health and job performance in a setting where these two outcomes are inextricably linked—the U.S. military. Our paper adds to the current understanding of how paid leave helps or hurts women's career trajectories by considering whether more generous paid maternity leave length is associated with women's performance and subsequent career advancement when they return to work.

2.4 Data

We draw data from the Marine's Total Force Data Warehouse and obtain records on all active-duty and reserve Marines who served at any point during January 2010 through December 2019. Our data include basic descriptive information on service member characteristics (age, gender, race/ethnicity, education status, and AFQT/GCT scores—which measure aptitude and intelligence), dependent characteristics for spouses and children (exact date of birth, gender, race/ethnicity, and whether a spouse is in the military), and job characteristics of the service member (job type, rank, time in service, time remaining in job commitment, and unit location). Our outcomes include two of the three primary measures

of job performance used for promotion and retention decisions: physical ability assessments and supervisor ratings of job proficiency. We do not have data on the third measure (rifle marksmanship assessments).

Our first outcome, physical performance, measures job proficiency among Marines based on standard physical fitness test scores. Scores are awarded on a 300-point scale, which is adjusted for age and gender such that women do not need to do as many push-ups as men, and older service members do not need to run as fast as younger ones to achieve the same score. We standardize raw physical fitness scores by year, gender, and test type (one in the 1st half of the year and the other in the 2nd). We combine the Z-scores for the two tests into one measure, generally observed twice per year per Marine. Due to the physically demanding nature of the assessment, women are not required to take the test when pregnant, and they are exempt from tests for 6 months after giving birth. We measure physical performance outcomes for mothers up until 10 months before they give birth and starting 8 months after. We resume measurement at 8 months postbirth due to concerns that commanding officers may allow some women whose test dates fall 7 months after birth to skip the test during that assessment round.

For our second outcome, supervisor ratings of job proficiency, we standardize scores by year, gender and test (ProCons vs. FitRep). We then combine the Z-scores of the two tests into one outcome we call supervisor performance ratings. Note, we are missing ProCons for junior enlisted who left the Marines before October 2017. However, we observe the full history of performance ratings (including ratings prior to October 2017) for any service member who was active as of October 2017. For our sample with complete ratings, we observe scores at least twice per year among junior enlisted, once in the 1st half and once in the 2nd half. For senior enlisted and officers, we observe supervisor ratings a minimum of once per

year. If a Marine is transferred, discharged, or promoted, or if their supervisor changes, they will receive additional performance ratings. Marines are relocated every few years, as are their designated supervisors. Decisions on relocation are made from a central location, which prevents Marines from manipulating their scores by selecting their supervisors (Cunha et al., 2018). Nonetheless, the subjective nature of these assessments means we cannot distinguish true changes in job performance from supervisors' *perceptions* of changes in performance using this measure.

Finally, using information on Marines' job rank over time, we can track promotion outcomes for Marines in our sample. We count the number of promotions a Marine receives relative to 10 months before they have a child. A value of 1 on the variable indicates 1 promotion achieved since the time point before pregnancy ($t = -10$).

Our preferred sample includes first-time parents who were active-duty 10 months before the birth, in other words right before the pregnancy began. To ensure our results are not driven by selective attrition, we require first-time parents remain in our sample for 12 months prior to the birth and 24 months after. We also include a group of "nonparents" in our sample, defined as Marines with at least 36 months total of service who do not experience a birth during the study window. Table 2.A.1 shows how the characteristics of the sample change based on a variety of possible sample restrictions.

We present characteristics of the first-time parents in our preferred sample in Table 2.1, alongside characteristics of first-time civilian parents who are employed and have a child under the age of 1. First-time parents in the Marines are younger than their civilian counterparts and have much lower rates of college attendance and college completion. First-time Marine mothers and fathers also identify as Black or Hispanic at higher rates than first-time civilian parents. In contrast, marriage rates are generally similar: 87% of Marine

vs. 83% of civilian fathers are married when they have their first child; and 68% of Marine vs. 78% of civilian mothers are married at first birth. In our sample, we rely on the Standard Occupational Classification (SOC) system, a federal standard used to classify workers into occupational categories, to explore the distribution of job types among Marines relative to civilians. We crosswalk Marine job codes to SOC codes and find that–outside of military-specific occupations–the largest share of first-time Marine fathers work in natural resources, construction, or maintenance (labeled "Construction/maint."), while the largest share of first-time Marine mothers work in sales or office roles.[8] Most civilians who have a first child and stay in the workforce tend to be employed in management, business, science, or arts.

Only a small share of first-time Marine parents in our sample are officers (akin to civilian managers): 15% and 8% of Marine mothers and fathers, respectively. As such, the vast majority of first births occur to enlisted Marines. Finally, new Marine parents score just above average on required intelligence tests, including the AFQT and GCT.

Based on descriptive differences between Marine and civilian first-time parents, results from our analyses may generalize best to younger and less-educated workers as well as workers of color.

2.5 Empirical Approach

The ideal experiment to isolate the causal effect of fertility on men and women's work performance would randomly assign pregnancy and parenthood to workers. Random assignment would ensure that–on average–differences in work performance were not driven by underlying characteristics of the types of men and women who chose to become parents but rather by

[8]Note that the vast majority of these jobs in the Marine Corps are office jobs but categorized under the umbrella of sales and office.

Table 2.1: Characteristics of First-Time Parents

	Fathers		Mothers	
	Marines	Civilians	Marines	Civilians
Age (mean)	25.5	31.6	23.5	29.9
Education				
Some College/Associates	5%	27%	5%	28%
Bachelor's Degree	17%	30%	10%	34%
Marital Status				
Married	87%	83%	68%	78%
Race/Ethnicity				
Black (Non-Hispanic)	9%	5%	15%	6%
Hispanic	14%	13%	21%	11%
Job Classification				
Mngmt./business/science/arts	10%	45%	14%	58%
Service	4%	11%	7%	15%
Sales/office	12%	15%	35%	24%
Construction/maint.	29%	15%	18%	0%
Production/moving/transpo.	14%	14%	19%	3%
Military Specific Chars.				
Officer	15%	–	8%	–
AFQT score (percentile)	63.2	–	58.7	–
GCT score (m=100; sd=20)	111.4	–	103.6	–
Observations	26,916	59,423	2,801	49,013

Notes: Displays characteristics of first-time parents in the Marine Corps in our sample alongside characteristics of first-time civilian parents in the labor market. Time-varying characteristics of Marines in our sample (e.g., age) are measured 10 months before the birth (t=-10). Data on civilians come from the American Community Survey 1-year estimates, 2010 to 2019. We limit the civilian sample to adults who are employed in the civilian labor market and have a first child under age 1. Job categories correspond to Standard Occupational Classification (SOC) system groups applied to U.S. Marine Corps job codes and available in the American Community Service. Military specific variables include whether a Marine is ranked as an officer (akin to manager) and AFQT and GCT scores, which are measures of intelligence. We do not observe these military-specific variables in the civilian sample.

the transition to parenthood itself. Of course, random assignment of childbirth/pregnancy to individuals in the workforce is both unethical and unfeasible. Yet, a simple post hoc comparison of parents relative to nonparents is unlikely to recover a causal estimate of the

effect of having a child. Those who opt into parenthood likely differ from nonparents in ways that might also correlate with work performance.

In the absence of a feasible experiment, we employ two alternative quasi-experimental approaches to estimating the impacts of parenthood. Broadly speaking, the first approach relies on within-person variation in the precise timing of pregnancy and birth as a shock to outcomes. In contrast, the second approach relies on between-person variation. We assign observably similar nonparents to placebo births and compare differences in outcomes for parents and nonparents across the birth and placebo birth events. For some outcomes, we are able to use both strategies to produce estimates. We show results are consistent across the approaches, increasing our confidence in each.

2.5.1 Impacts of Parenthood on Job Performance

To study parents' health-related job performance outcomes, our primary identification strategy exploits variation in the precise timing of births to identify the effect of childbirth on first-time parents. If the transition to parenthood has an impact on health and performance, then the birth should generate a sharp change in these outcomes directly after it occurs. We can attribute any discontinuity in the outcomes at the time of the birth to the birth itself if we assume that other factors that shape job performance do not also undergo a sharp change in the same month as childbirth. In other words, while the choice to have a child may be endogenous, the exact timing of conception and subsequent childbirth serves an exogenous shock to the outcomes of interest. We consider this assumption reasonable, given that we rely on monthly data. We can separately identify the impact of childbirth from anything else that might occur around the time of pregnancy or birth but outside the birth month.

We build in a second source of variation to our event study approach by including Marines

who do not give birth during the study window in the analysis. In this way, our event study is a form of a difference-in-differences model. nonparents in the analysis help approximate counterfactual trends in outcomes that all Marines would have experienced, assuming outcomes would have evolved similarly between first-time parents and nonparents absent childbirth. We test whether parents' and nonparents' outcomes evolve similarly during the prepregnancy period and find reasonable confirmation of parallel trends. We also include nonparents in the event study analysis to limit the share of 2 x 2 comparisons in the event study estimation that use early-treated units (i.e. parents with first births early in the study window) as controls. In this way, we avoid bias in our estimates that can result from staggered treatment-timing event study approaches (see Baker et al. (2021) for an overview).

To implement our event study approach, we begin by estimating a fully flexible, dynamic specification separately for men and women, which allows us to test the assumption that outcomes change smoothly prior to the pregnancy and similarly for both parents and nonparents in that period. Our model is of the following form:

$$Y_{it} = \alpha_i + \phi_t + \sum_{r=k_{min}}^{k_{max}} \mathbb{1}(t = t_i^* + r)\beta_r + \varepsilon_{it} \tag{2.1}$$

where β_r represents the effect of a birth in month t_i^* on outcomes r months later (or r months before, if $r < 0$). Effects are measured relative to month $r = -10$, which corresponds to 10 months prior to the birth and approximately 1 month before the start of the pregnancy. In other words, β_{12} would represent the average outcome 12 months after the birth, relative to $r = -10$ (the month before pregnancy). We censor r at $k_{min} = -18$, or $k_{min} = -24$, depending on the outcome.[9] We include α_i, an individual fixed effect to account for stable

[9]We examine physical performance scores 24 months prior to childbirth but measure supervisor ratings of job performance only 18 months prior to childbirth. Supervisor ratings cover roughly a 6-month retrospective period; therefore, any rating awarded 18 months before the birth implicitly covers an earlier time period of

individual differences; ϕ_t, a month-by-year fixed effect to account for general changes over time in the outcome (e.g., due to changes in fitness test standards in a particular year); and ε_{it} is the error term. As discussed, the estimation of month-by-year fixed effects is assisted by the inclusion of nonparents in the data, who provide an estimate of universal time-patterned changes to the outcome that affect all Marines similarly. We run all models separately by gender.

Beyond estimating individual month coefficients β_r, our goal is to identify any declines (or improvements) in performance during pregnancy, any drops immediately following birth, and any recovery following the immediate impact of birth. Similar to Lafortune et al. (2018), we create a more parsimonious model of performance changes over time relative to before the pregnancy by defining:

$M_{it}^{pregnancy} = t - (t_i^* - 9)$, if person i has a baby at time t_i^* and $t_i^* - 9 \leqslant t < t_i^*$, and $M_{it}^{pregnancy} = 0$ otherwise (for monthly trends during pregnancy);

$M_{it}^{drop} = 1$, if person i has a baby at time t_i^* and $t > t_i^*$ and $M_{it}^{drop} = 0$ otherwise (for postnatal drops following birth);

$M_{it}^{recovery} = t - (t_i^* + q)$, if person i has a baby at time t_i^* and $t > t_i^* + q$, where q is time point when the person is eligible to be observed for the given outcome after the birth, and $M_i t^{recovery} = 0$ otherwise (for monthly trends above and beyond the drop in level); and

$M_{it}^{\Delta recovery} = t - (t_i^* + 12)$, if person i has a baby at time t_i^* and $t > t_i^* + 12$, and $M_{it}^{\Delta recovery} = 0$ otherwise (for any change to the monthly recovery rate that begins at 13 months).

Using these different time frames, we then estimate a semi-dynamic specification that fits

performance.

linear splines to portions of the data, as follows:

$$Y_{it} = \alpha_i + \phi_t + M_{it}^{pregnancy}\beta_1 + M_{it}^{drop}\beta_2 + M_{it}^{recovery}\beta_3 + M_{it}^{\Delta recovery}\beta_4 + \varepsilon_{it} \tag{2.2}$$

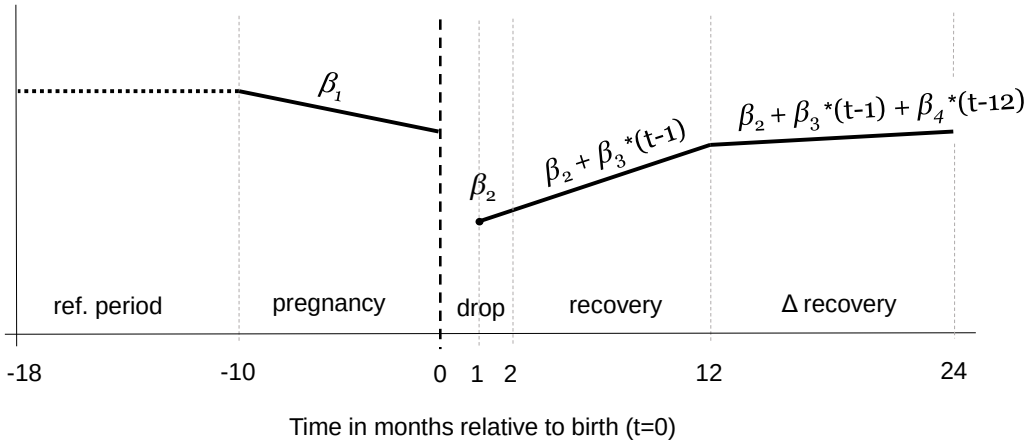
Here, β_1 captures the monthly linear change in the outcome during the pregnancy period ($t = [-9, -1]$), relative to the prepregnancy period average ($t \leqslant -10$). The effect captured by β_2 represents the acute postnatal drop (if any) in the outcome in the 1st month parents are again assessed after childbirth. Then, β_3 captures the monthly linear recovery in the outcome following that initial drop, and β_4 captures any change in the monthly linear recovery rate in the child's 2nd year of life ($t = [13, 24]$). All parameters are measured relative to the prepregnancy period average ($t \leqslant -10$). We present a diagram of this model in Figure 2.1. We use this semi-dynamic spline specification as our main model to estimate the magnitude of and confidence internal around the impact of birth on health and job performance outcomes. We present parameters from this model and their implications for effects at 12 and 24 months postbirth in subsequent tables of results.

2.5.2 Impacts of Parenthood on Career Advancement

Beyond changes in job performance, we explore whether having a child impacts men and women's career advancement. Unlike with our job performance measures, we cannot study promotions using an event study framework. Promotions do not occur often, one's propensity to be promoted is also a function of how recently they received their last promotion. As a result, to investigate differences in promotion outcomes due to having a child, we follow Kleven et al. (2019b) by assigning a comparison group of nonparents to placebo births based on observable characteristics that predict selection into parenthood. We then compare

Figure 2.1: Stylized Representation of the Semi-Dynamic Specification, Equation 2.2:

$$Y_{it} = \alpha_i + \phi_t + M_{it}^{pregnancy}\beta_1 + M_{it}^{drop}\beta_2 + M_{it}^{recovery}\beta_3 + M_{it}^{\Delta recovery}\beta_4 + \varepsilon_{it}$$



Notes: Figure displays a diagram of parameters defined in Equation 2.2 where the postbirth drop (β_2) is estimated in the 1st month following pregnancy. This holds for all models among the sample of men. Among the sample of women, we begin measuring the postbirth drop (β_2) for physical fitness performance at 8 months and supervisor ratings at 6 months after the birth. We also cannot estimate β_1, the pregnancy trend, in the model on women's physical fitness outcomes, due to restrictions on when pregnant women's fitness is assessed.

promotion outcomes across the "birth event" for Marines with actual births and those with placebo births to trace whether gaps in outcomes arise.

Our placebo birth assignment process relies on an adaptive ridge LASSO (least absolute shrinkage and selection operator) model with 10-fold validation to determine the best predictors of a first birth. Possible predictors include age, race/ethnicity, military entrance exam scores (AFQT scores), marital status (including whether a spouse is also in the military), level of education, occupational field, and most recent physical performance score, as well as interactions among all variables. We run this predictive model separately among women and men, measuring all characteristics of first-time parents at $t = -10$, 10 months before

the birth. Second, we use the predicted propensity from step one to match parents to up to five most observably similar nonparents. We require matches to have the same job rank, the same number of months in service, and to have enlisted in the Marines in the same calendar year. We then assign placebo births to nonparents 10 months after the time of the match (at $t = -10$). The exact match on job rank, number of months in service, and year of entry ensures that Marines with and without children have similar promotion histories and work contexts before the pregnancy begins.

Analyses then compare the average changes in outcomes for first-time parents to the average change in outcomes for the 5 most observably similar nonparents to whom they match. Each parent receives a weight of 1 in the analysis, while each match receives a weight of 0.2 per match-month in the case where 5 distinct nonparents match to each parent.[10] We conduct the match at $t = -10$. However, we estimate differences in outcomes between first-time parents and nonparents *before* $t = -10$, as well. If the nonparents provide a suitable comparison, we expect their outcomes will look similar to the parents' in the months before the pregnancy ($t < -10$).

To lend additional confidence to this research design, we also use it to reestimate the impact of childbirth on Marines' physical fitness performance and supervisor ratings. We observe consistent findings across the event study and placebo birth strategies, increasing our confidence that the placebo birth approach generates plausible causal estimates of the impact of parenthood on promotion and vice versa.

[10]We conduct nearest-neighbor matching with replacement, meaning parents can match to less than 5 nonparents who then get higher weights. Additionally, the same nonparent can be matched to different first-time parents.

2.5.3 Maternity Leave Policy Impacts

Finally, we investigate variation in the impact of parenthood on mothers' job performance and career advancement across maternity leave policies.[11] Naïve evaluations of parental leave policies may not be causal because in most settings the quantity of leave is not randomly assigned. In many settings, for instance, certain types of mothers (e.g., more advantaged mothers) may make choices that afford them more leave, such as opting to work at firms that provide more leave, but these mothers may have higher (or lower) work performance for reasons unrelated to the amount of parental leave provided. To interpret policy impact estimates as causal, it must be the case that women's potential job performance outcomes are not correlated with the length of their maternity leave. As discussed, some changes to leave length in our setting were unexpected and applied to women who were already pregnant, while other changes provided some advance notice. We explore empirical evidence regarding selection into maternity leave length and present findings in our results section.

We group first-time mothers into four policy categories, P_{it}^{j}, based on the amount of paid leave available at the time of birth. We define P_{it}^{1} as the initial 6-week policy (baseline); P_{it}^{2} as the retroactive 18-week policy, where parents expected 6 weeks of leave but actually received an additional 12 weeks of leave (that could be used discontinuously) once back at work; P_{it}^{3} as the 18-week policy where mothers had the option to take 18 weeks of leave continuously and largely did (Bacolod et al., 2020); and P_{it}^{4} as the policy that afforded 12

[11] We considered regression discontinuity (RD) as an alternative strategy for the leave analysis, given that unexpected policy changes apply to births on either side of precise birth-date cut points. However, the cut points happen to fall near the end of the fiscal and calendar year—which may affect outcomes for other reasons. To address this, one option would be to embed an RD approach within a difference-in-differences strategy, comparing outcomes 3 months before and after the unexpected policy change in the year of the reform to years where no reform occurs for example, following Persson and Rossin-Slater (2019). However, we are concerned about statistical power given the small number of women in our dataset.

weeks of continuous paid leave. We then estimate:

$$\begin{aligned} Y_{itmy} =& \alpha_i + \gamma_t + P_{it}^{j}(M_{it}^{pregnancy}\beta_{j1} + M_{it}^{drop}\beta_{2j} + M_{it}^{recovery}\beta_{3j} \\ &+ M_{it}^{\Delta recovery}\beta_{4j}) + \varepsilon_{itmy} \end{aligned} \tag{2.3}$$

where each coefficient represents the pregnancy trend β_{1j}, postnatal drop β_{2j}, recovery β_{3j}, and change to recovery β_{4j} for each policy. We use the same policy subgroups, P_{it}^{j}, to compare subgroups of effects on promotion trajectories among matched parents and nonparents. We group women into policy categories based on the timing of births/placebo births.

Note, we group women into P_{it}^{3}, so long as their birth occurred during the 18-week policy time window, regardless of whether they learned of extended leave when already pregnant or became pregnant after the extended leave policy was announced. We compare differences in observable characteristics across these two groups and find little evidence of selection into the 18-week policy among women who became pregnant after the policy was announced (see Table 2.A.2). We also do not find evidence that births increase (or decrease) 9 months following the policy announcement, which might suggest selection into (or out of) parenthood induced by longer leave. Figure 2.A.1 shows the density of births across April 2016 (9 months after the policy announcement); there is no statistically significant difference across this threshold.

2.6 Results

2.6.1 Main Impacts

We begin by examining evidence on our identifying assumptions, namely that outcomes leading into the pregnancy and birth evolve smoothly and that nonparents provide a suitable

counterfactual estimate of general time trends in outcomes.

Figure 2.2 presents results from our flexible event study model estimated using Eq. 2.1 with point estimates for prepregnancy effects displayed in shaded blue on each graph. We conduct F-tests to assess whether the individual month coefficients β_r from Eq. 2.1 jointly differ from zero for months $r \leqslant -10$. The significance value for each F-test is presented in Figure 2.2 below each graph. We also present the slope parameter for the pretrend estimate alongside the standard error and significance of the estimated slope.[12] Overall, the parallel trends assumption appears to be reasonably satisfied, especially in our sample of mothers. Among fathers, some evidence suggests that physical performance scores and supervisor ratings are trending positively before the pregnancy. We interpret our findings on fathers by taking into account these patterns.[13]

The event study estimates of the impact of parenthood displayed in Figure 2.2 Panel A show women and men perform more poorly on job-related physical performance assessments when initially assessed after having a child. For women, performance declines are large and persistent. Even 24 months after giving birth, women's physical assessment scores continue to be lower than they were before the pregnancy. For men, performance declines begin during the mother's pregnancy and reach their lowest point 1 month after the child's birth. Declines are short-lived. By 12 months after the birth, men appear to perform at their prepregnancy

[12]We define $M_{it}^{pretrend} = t - (t_i^* - 10)$ if person i has a baby at time t_i^* and $t < t_i^* - 9$ (and 0 otherwise); we then estimate:

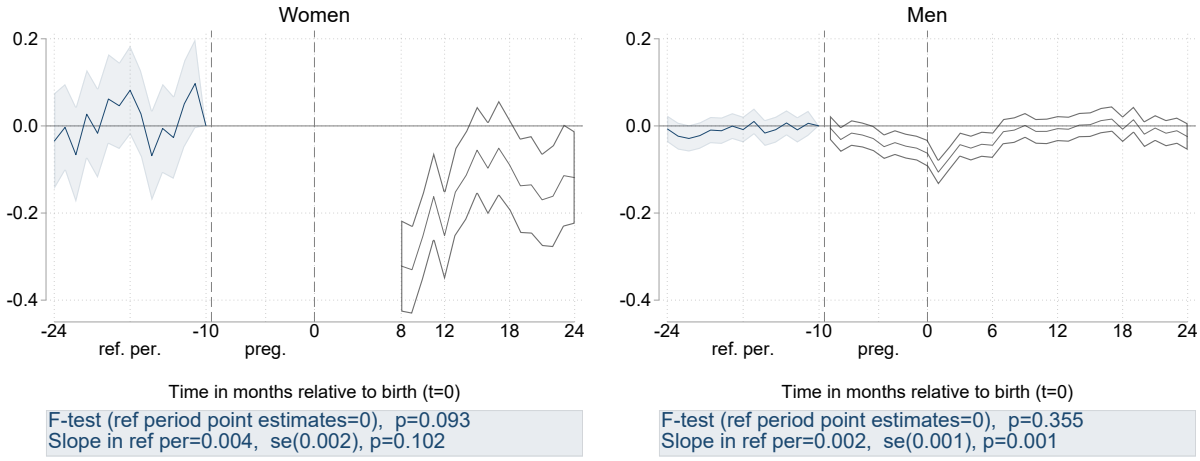
$$Y_{it} = \alpha_i + \phi_t + X_{it}\theta + M_{it}^{pretrend}\beta_1 + M_{it}^{pregnancy}\beta_2 + M_{it}^{drop}\beta_3 + M_{it}^{recovery}\beta_4 + M_{it}^{\Delta recovery}\beta_5 + \varepsilon_{it} \quad (2.4)$$

This matches Equation 2.2 except that it adds a linear trend for all of the prepregnancy months other than $r = -10$, which serves as the reference. We expect the event of pregnancy/birth to drive performance differences only *after* the pregnancy has taken place. Therefore, we expect to find no evidence of differences in trends before the pregnancy.

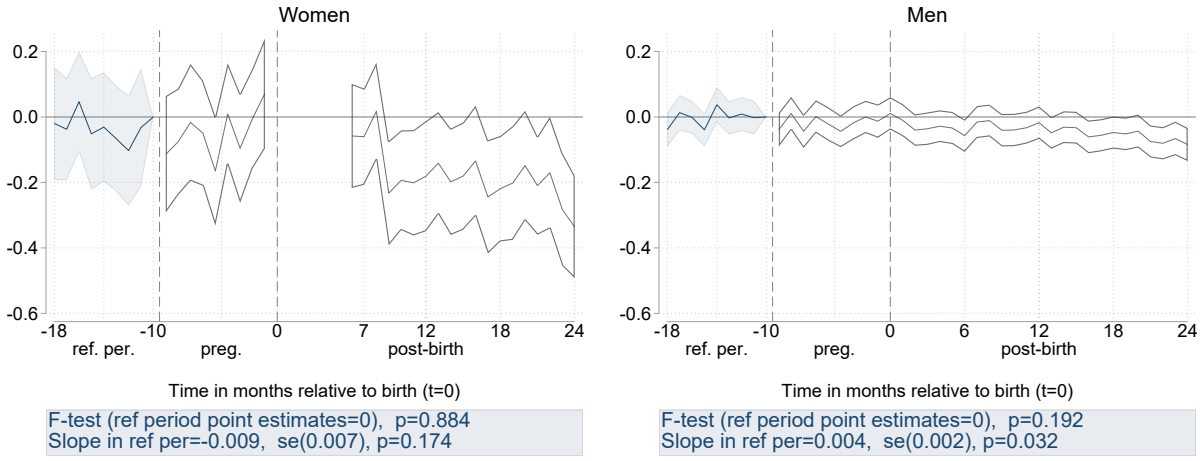
[13]Sharp discontinuities after the birth in fathers' physical performance scores suggest effects are not merely a continuation of preexisting pretrends or level differences for this outcome. For supervisor ratings, despite some evidence of pretrends among fathers before the pregnancy, we do not see large deviations in outcomes after the birth.

Figure 2.2: Event Study Estimates of the Impact of Birth on Job Performance

(a) Physical Performance Scores



(b) Supervisor Ratings



Notes: Displays coefficients from event study regressions. Outcomes include (1) standardized scores (mean=0, SD=1) from physical/combat fitness tests and (2) standardized (mean=0, SD=1) scores from supervisor-rated job performance evaluations, measured separately for males and females. Sample includes first-time parents observed at least 12 months before and 24 months after birth, as well as a control group of same-gender Marines who do not have a birth during the study window and remain in service for at least 36 months. Regressions include fixed effects for individuals and months by year. The reference month is 10 months before the birth ($t = -10$), capturing the time point before the start of pregnancy. Vertical dotted lines reflect the start of the pregnancy ($t = -9.5$) and the birth (t=0). Standard errors are clustered at the individual level and included as shaded/hollowed areas representing a 95% confidence interval.

levels on physical fitness tests. Figure 2.2 Panel B shows evidence of relatively consistent lower supervisor ratings for women in the 2 years after having a child, though estimates are noisier. There appears to be minimal to no impact of having a child on fathers' supervisor-rated job performance.

To quantify the size of the impacts and draw statistical inference, we next estimate a more parsimonious, semi-dynamic parametric model that corresponds to patterns in our flexible models (diagrammed in Figure 2.1). The results from the semi-dynamic specification are summarized in Table 2.2. The parameters in Table 2.2 include: a pregnancy trend, which captures any change in the outcome from month to month during the pregnancy; a postbirth drop, which captures any level shift in the outcome during the 24-month postbirth period; a recovery trend, which captures any monthly changes in the outcome during the 24-month postbirth period; and a "Δ recovery" trend, which reflects any change in the recovery slope during the 13–24 postbirth period. All parameters are measured relative to the prepregnancy period. Recall, women are not subject to physical fitness tests while pregnant, so we do not estimate a pregnancy trend for women's physical performance. After the birth, we measure women's physical performance scores starting 8 months postbirth due to exemptions from fitness testing related to pregnancy. We measure women's supervisor ratings starting 6 months postbirth to account for maternity leave. Also, recall that to remain in the sample, we require individuals to remain "on the job" as Marines for at least 24 months following the birth.

Mothers' Physical Fitness Scores

Columns (1) and (2) in Table 2.2 present results of the impact of parenthood on women's physical fitness performance scores and supervisor-evaluated job proficiency ratings. Begin-

Table 2.2: Impacts of Childbirth on Job Performance Among First-Time Parents

	Women		Men	
	Physical Performance (1)	Supervisor Ratings (2)	Physical Performance (3)	Supervisor Ratings (4)
Model Parameters				
Pregnancy trend	–	0.006	-0.004***	0.004*
		(0.005)	(0.001)	(0.001)
Postbirth drop ($birth - 24mos.$)	-0.362***	0.044	-0.067***	0.020
	(0.031)	(0.041)	(0.006)	(0.011)
Recovery trend ($birth - 24mos.$)	0.055***	-0.018*	0.009***	0.002
	(0.010)	(0.008)	(0.001)	(0.001)
Δ Recovery trend ($13 - 24mos.$)	-0.053***	0.014	-0.011***	-0.005*
	(0.012)	(0.010)	(0.001)	(0.002)
Estimated Effects				
12-month effect	-0.142***	-0.064	0.027***	0.042***
p-value	[0.000]	[0.099]	[0.000]	[0.000]
24-month effect	-0.116***	-0.114**	0.001	0.009
p-value	[0.000]	[0.009]	[0.857]	[0.466]
Prepregnancy mean	0.140	-0.015	0.295	0.011
N of Individuals	20,470	9,854	270,630	112,968
Observations	154,070	83,385	2,066,435	986,863
R^2	0.587	0.412	0.594	0.422

Notes: Displays coefficients from Eq.2.2, the semi-dynamic event study specification. Outcomes include (1) standardized (mean=0, SD=1) scores from physical/combat fitness tests conducted 2x per year and (2) standardized scores (mean=0, SD=1) from supervisor-rated job performance evaluations conducted 1-2x per year. Women's supervisors ratings are not measured 0 to 5 months postbirth due to overlap with maternity leave. Women's physical performance scores are not measured 9 months before through 7 months after birth because women are not required to take fitness tests during and after pregnancy. Regressions include individual and month-by-year fixed effects. The parameter "Pregnancy trend" captures trends during pregnancy, if observed. "postbirth drop" is an indicator equal to 1 after the birth, starting in $t = 1$ for all men's outcomes; $t = 8$ for women's physical performance; and $t = 6$ for women's supervisor ratings. "Recovery trend" estimates monthly changes in the outcome for the entire postbirth period. "Δ Recovery trend" estimates any change in the slope in the second year postbirth. Robust standard errors are clustered by individual, shown in parentheses. p-values that test whether 12-month and 24-month average effects differ from zero are shown in brackets. *** $p < 0.001$, ** $p < 0.01$, $^{*}p < 0.05$..

ning at 8 months postbirth when women are generally required to take fitness tests again, mothers perform 0.36 standard deviations below their prepregnancy average. Mothers recover from this initial drop at a rate of about 0.06 standard deviations per month in the 1st year of the child's life. By 12 months postbirth, women's expected physical performance scores are 0.14 standard deviations below their prepregnancy levels (p-value[12-month effect]=0.000). Mothers' physical fitness recovery slows to nearly zero in the child's 2nd year of life (recovery during 13–24 months postbirth is the combination of the main recovery trend of 0.06 and the additional Δ recovery trend of –0.05). Two years after having a baby, mothers' predicted physical performance remains 0.12 standard deviations lower than before the pregnancy (p-value[24-month effect]=0.000).

Recall, physical performance assesses a combination of cardiovascular health, endurance, and upper-body and core strength. Table 2.A presents raw scores for each item on the fitness assessments that measure overall physical performance. Declines in women's performance occur across almost all measured assessments. When women return to testing, the initial "postbirth drop" shows women run more slowly (i.e., run times increase) and complete fewer crunches and overhead lifts (i.e., crunch and overhead lift counts decrease). Note, before 2017, women could opt into or out of the pull-up assessment. Likely, as as a result, overall patterns diverge for this outcome, and we interpret these results with caution.

Mothers' Supervisor Ratings

Unlike the impact of childbirth on women's physical performance, its impact on supervisor job proficiency ratings gradually accrues over time. Supervisor ratings do not change during pregnancy nor 6 months after the birth when we again measure ratings for mothers (indicated by the pregnancy trend and postbirth intercept in Table 2.2, column (2)). Instead, mothers'

supervisor ratings steadily decline during the postbirth period at a small but statistically significant rate of –0.02 standard deviations per month. Two years after becoming a parent, women are rated 0.11 standard deviations below their prepregnancy levels of job performance (p-value[24-month effect]=0.009).[14]

Fathers' Physical Fitness Scores

Next, we turn to outcomes among men who become fathers, shown in Table 2.2, columns (3) and (4). The pattern of effects on fathers' job-relevant physical performance is consistent with mothers' but smaller in magnitude. Marine men's physical fitness scores decrease at a rate of 0.004 standard deviations per month during the mother's pregnancy. There is a substantial drop in physical performance in the month immediately following the birth to 0.07 standard deviations below prepregnancy performance levels. Fathers on average recover by the child's first birthday, such that new fathers' physical fitness actually exceeds their prepregnancy levels by 0.03 standard deviations 12 months after their child's birth (p-value[12-month effect]= 0.000). The rate of recovery in physical fitness performance flattens for fathers in the 2nd year postbirth (reflected by adding the recovery trend to the Δ recovery trend). Twenty-four months after having a child, fathers' physical ability scores are no higher than they were before the pregnancy. Though we observe small statistically significant improvements in fathers' physical performance at 12 months after the birth, we interpret this cautiously given the small positive pretrends observed before the pregnancy (see Figure 2.2).

Table 2.A shows changes in fathers' raw, item-level performance on fitness tests. Initial

[14]Though supervisors are trained for consistency on job performance evaluations, we cannot distinguish true differences in performance from supervisors' *perceptions* of differences in performance among workers after they have a child.

declines in performance appear to come from slower run times (an increase in time in seconds) and fewer pull-ups (a decrease in number) immediately after the birth. While men appear to do *more* crunches after having a child and perform slightly better on each measured item 12- and 24-months after the birth, we note these impacts are small in magnitude and interpret them with caution due to evidence of small pretrends in outcomes for men before the birth.

Fathers' Supervisor Ratings

The impacts of the transition to fatherhood on men's supervisor-rated work performance in the overall sample are minimal, and again, we interpret findings with caution given evidence of small pretrends in the outcome before pregnancy. Table 2.2, column (4) shows small positive trends in men's supervisor ratings during the mother's pregnancy, with ratings increasing by 0.004 standard deviations per month up until childbirth. The postbirth parameter and recovery trend are not statistically significant. However, when combined to estimate a cumulative 12-month effect, fathers' supervisor performance ratings are 0.04 standard deviations higher than before the mother's pregnancy (p-value=0.000). During the child's second year of life (months 13 to 24), men's performance declines at a rate of –0.005, such that by 24 months after the birth, there is no predicted change in supervisor ratings among fathers relative to before the birth. The small magnitude of these estimated effects, combined with a pattern of largely null results shown in Figure 2.2, leads us to conclude there is a negligible impact of parenthood on fathers' job proficiency ratings.

Robustness of Event Study Impacts

Tables 2.A.4 and 2.A.5 show the sensitivity of our results to various sample restrictions. Across columns (1) to (5) in each table, we progressively restrict the sample to minimize

attrition over the study window. This allows us to explore the extent to which our findings are driven by compositional differences in terms of who remains in the Marines. Column (1) in Tables 2.A.4 and 2.A.5 presents estimates for the full sample of first-time parents and nonparents, placing no restrictions on the length of time they must remain in our study window to contribute to the estimates. Column (2) restricts first-time parents to those observed at least 12 months before and 12 months after the birth, and column (3) adds that nonparents serve a minimum of 24 months to reflect the restriction placed on first-time parents in column (2). In columns (4) and (5), we restrict to first-time parents observed at least 12 months before and 24 after the birth. Column (4) places no restriction on nonparents, and column (5) requires nonparents included in the sample to have remained in service for at least 36 months. Column (5) is our preferred sample and is used to estimate the main impacts shown in Table 2.2.

Among women, the impact of childbirth on women's physical fitness and supervisor-rated performance is largely consistent across all five samples shown in Table 2.A.4. When we place fewer restrictions on the sample, we detect statistical significant negative impacts of birth on supervisor ratings as soon as 12 months postbirth (see Panel B, columns (1)–(3)). When we restrict the sample to women who remain in service for at least 24 months after the birth, we no longer detect statistically significant declines in supervisor ratings 12 months postbirth, but the magnitude and direction of the estimates remains similar. If anything, then, this pattern of findings lends confidence to our finding that women's supervisor-rated job performance declines after having a child.

Among men, Table 2.A.5 shows that the direction of effects of childbirth on fathers' physical performance is generally consistent across samples. However, the magnitudes of the coefficients differ slightly, such that for less restrictive samples (columns (1)–(3)), fathers

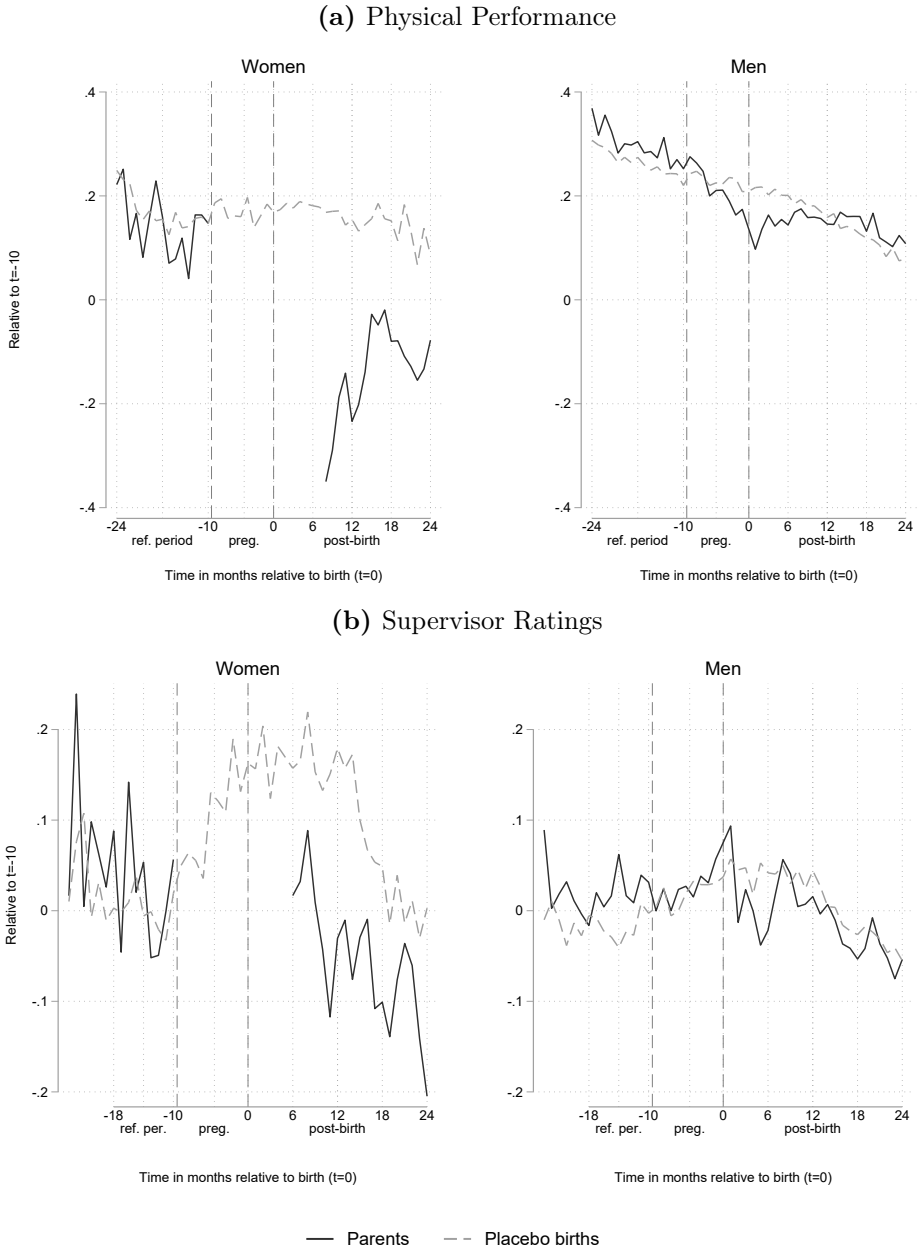
perform, on net, worse in terms of physical performance 12 months and 24 months after having a child. These negative impacts are small, ranging from 0.04 to 0.03 standard deviations, and dissipate once we require first-time fathers remain in the sample for longer, through 24 months after the birth. This pattern suggests that new fathers who do not stay in the sample past 12 months after having a child (those included in specifications (1)–(3) but omitted from specifications (4)–(5)) may be more negatively affected in terms of physical fitness after having a child. This same pattern across samples holds for impacts on fathers' supervisor ratings but to a smaller degree.

Mothers' & Fathers' Career Advancement

Before we turn to the impacts of childbirth on first-time parents' career advancement, we first replicate our results on supervisor ratings and physical performance using the matching design that assigns placebo births to nonparents. Figure 2.3 shows patterns of job performance outcomes between parents and nonparents assigned placebo births. Gaps in outcomes between parents and nonparents emerge only after childbirth for both men and women. Results are generally consistent with findings from our event study models, suggesting the placebo birth strategy is a reasonable approximation of our event study design.

Figure 2.4 presents promotion trajectories of parents and nonparents across childbirth, separately for fathers and mothers. We measure the outcome as the number of promotions relative to $t = -10$, the time period before the pregnancy began. Among fathers and nonfathers, promotion trajectories evolve nearly identically. By 24 months after having a child, fathers on average have achieved one additional promotion beyond their job rank as measured in the month prior to the pregnancy, as have nonfathers. In contrast, gaps in career advancement emerge between mothers and nonmothers starting during women's

Figure 2.3: Placebo Birth DiD Estimates of the Impact of Birth on Job Performance



Notes: Displays differences in outcomes between first-time parents and nonparents across birth/placebo birth events for each month before and after birth. Outcomes include (1) standardized (mean=0, SD=1) scores from physical/combat fitness tests conducted 2x per year and (2) standardized scores (mean=0, SD=1) from supervisor-rated job performance evaluations conducted 1-2x per year. Matching occurs at $t = -10$, the time point before the start of pregnancy. Parents are matched to up to five most observably similar nonparents based on age, race/ethnicity, military entrance exam scores (AFQT scores), marital status (including whether a spouse is also in the military), level of education, occupational field, and prior physical performance scores. nonparents and parents must match exactly on job rank and number of months in service at $t = -10$. Vertical dotted lines reflect the start of the pregnancy ($t = -9.5$) and the birth ($t = 0$).

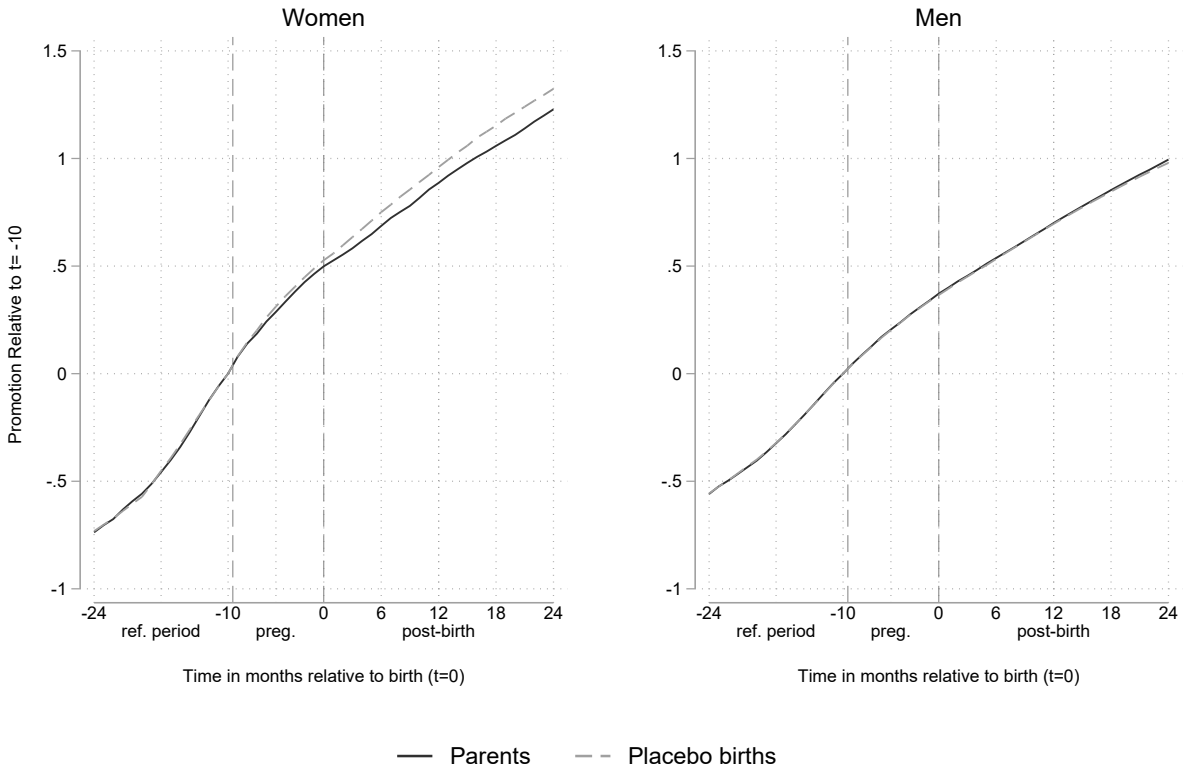
pregnancies. These gaps continue to widen through 24 months postbirth, such that first-time mothers have received 0.1 fewer promotions than counterfactual nonmothers, relative to their job rank before the birth. Importantly, trends in mothers and nonmothers' outcomes overlapped before the pregnancy began, indicating that the two groups were on similar career trajectories prior to the pregnancy.

2.6.2 Variation in Impacts by Maternity Leave Length

So far, we have presented aggregate impacts of birth on job outcomes among men and women. We now disaggregate these impacts according to the length of maternity leave in place when a woman gave birth. The key question of interest is whether longer or shorter leave predicts better or worse performance outcomes when women return to work.

As discussed, variation in leave length is, at times, quasi-randomly assigned; some policy changes were unexpected and applied to women who were already pregnant, while other changes were prospective, allowing women to potentially select in. Selection presents the biggest concern for women who gave birth under the 12-week policy, given that the reduction in leave length was announced prospectively. Table 2.A.6 presents descriptive characteristics of the women who gave birth under the four leave policy groups we analyze. We find minimal evidence of selection into leave length based on observable characteristics. Exceptions include significant differences in mothers' age at birth (though these are very small), percent who identify as Hispanic, percent in office jobs, and scores on the AFQT entrance test. When we estimate variation across policy groups, using the matched sample of parents and nonparents assigned to placebo births, our matching strategy implicitly controls for these underlying differences. We also test whether the density of first births is continuous across each policy threshold, providing additional evidence on possible selection.

Figure 2.4: Placebo Birth DiD Estimates of the Impact of Birth on Promotion Trajectories



Notes: Displays differences in the number of promotions relative to $t = -10$ between first-time parents and nonparents across birth/placebo birth events for each month before and after birth. nonparents assigned to placebo births are limited to those whose rank and number of months in service is an exact match with parents' 10 months before the birth. Among those with an exact match on rank and months in service, each parent's outcomes are compared to the five nonparents most similar to parents in their propensity to have a child based on age, race/ethnicity, military entrance exam scores (AFQT scores), marital status (including whether a spouse is also in the military), level of education, occupational field, and physical performance scores. The reference month is 10 months before the birth ($t = -10$), capturing the time point before the start of pregnancy. Vertical dotted lines reflect the start of the pregnancy ($t = -9.5$) and the birth ($t = 0$).

Figure 2.A.2 shows month-by-month variation in the density of births, including a test for any discontinuity across policy thresholds, following Cattaneo et al. (2018). None of the differences across the policy thresholds reach statistical significance, suggesting the policies did not influence female fertility itself.

Leave Length & Women's Physical Performance

Using the policy groupings described, we turn to variation in mothers' job-related physical fitness scores by considering the length of mothers' maternity leave after having a baby. Results are presented in Table 2.3, Panel A. We include the overall estimated effect as the first row of the panel. The bottom section of the panel shows estimates of effects by the length of leave.

Among mothers across all maternity leave policies, physical performance drops when initially observed again 8 months after having a baby. All of the drops measured 8 months after the birth statistically differ from zero. However, women who had longer maternity leave have larger physical performance declines when measured 8 months after having a child. Moreover, an F-test indicates that the magnitude of the drops across leave policies statistically differ from each other (p-(diff)=0.011). Put differently, we find evidence that the degree of decline in performance is statistically different based on the length of maternity leave mothers received. We also test whether the size of the effects 8 months after the birth vary among the women who had longer than 6 weeks of leave (i.e., 6 + 12 flexible weeks vs. 18 weeks vs. 12 weeks). We find marginally statistically significant differences (displayed in the row labeled "p(diff), >6 weeks of leave" = 0.081).

Results regarding physical performance measured 12 months after birth show that mothers who were given a retroactive 12 weeks of leave after they had given birth recover by the

Table 2.3: Impact of Childbirth on Women's Job Outcomes, by Maternity Leave Policy

A. Physical Performance (Event Study Estimates)						
	8 months postbirth		12 months postbirth		24 months postbirth	
	Effect size	p	Effect size	p	Effect size	p
Main effect:	−0.359***	0.000	−0.137***	0.000	−0.110***	0.000
Effects by leave length:						
6 weeks	−0.287***	0.000	−0.114***	0.000	−0.034	0.332
6 + 12 flex weeks	−0.713***	0.000	0.006	0.948	−0.392***	0.001
18 weeks	−0.470***	0.000	−0.178***	0.000	−0.145*	0.010
12 weeks	−0.335***	0.000	−0.198***	0.001	−0.256***	0.000
F-test of differences:						
p(diff), all effects	0.011		0.279		0.001	
p(diff), >6 weeks of leave	0.081		0.167		0.119	

B. Promotion Counts (Matched Sample Estimates)						
	1 month postbirth		12 months postbirth		24 months postbirth	
	Effect size	p	Effect size	p	Effect size	p
Main effect:	−0.035*	0.026	−0.076***	0.000	−0.105***	0.000
Effects by paid leave length:						
6 weeks	−0.058**	0.003	−0.135***	0.000	−0.132***	0.000
6 weeks + 12 flex	−0.004	0.934	−0.122	0.093	−0.128	0.110
18 weeks	−0.043	0.088	−0.040	0.240	−0.109**	0.006
12 weeks	0.031	0.249	0.053	0.188	−0.033	0.462
F-test of effects:						
p(diff), all effects	0.019		0.000		0.007	
p(diff), >6 weeks of leave	0.084		0.055		0.350	

Notes: Displays the expected value of the outcome across 3 time points after childbirth. Estimates in Panel (a) that come from Eq.2.3, the semi-dynamic event-study specification, interacted with indicators for paid maternity leave length. Estimates in Panel (b) that come from the comparison of parents and matched non-parents, interacted with indicators for paid maternity leave length. Policy groups include the initial 6-week leave allowance, the retroactive 18-week policy (6 weeks of leave plus an additional 12 weeks once back at work to be used discontinuously), the 18-week policy, and the 12-week policy. "p(diff), all effects" presents the p-value for an F-test of differences in the estimates across all policy periods. "p(diff), | > 6 weeks of leave" presents the p-value for an F-test of differences in the estimates among policy periods with greater than 6 weeks of leave. Regressions (including the main effects presented here) exclude women with births in November and December 2016 because women with births in these months could have fallen into one of two policy periods based on their doctor-estimated date of conception, which we do not observe. Robust standard errors are clustered at the individual level in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

child's 1st birthday. Notably, these mothers were only eligible to use the additional 12 weeks of leave before the child's 1st birthday. Bacolod et al. (2020) show that mothers used this leave as flexible time off during the child's 1st year. It is possible that, for these women, the flexible time off assisted with performance recovery. Women under all other leave policies continued to show physical performance declines 12 months after birth. However, the F-test of the differences in the magnitude of impacts across leave policies at 12 months postbirth is not statistically significant.

By the child's 2nd birthday (24 months after the birth), most women continue to score lower on employer-assessed physical fitness measures. However, by this point, women who received the shortest time off (6 weeks) perform no differently from their prepregnancy level of fitness. The F-test of differences in magnitude of the coefficients across all policy groups at 24 months after birth is statistically significant (p(diff)=0.001). There is no statistical evidence that the size of performance impacts differ among the policies longer than 6 weeks (p(diff), >6 weeks of leave=0.119).

Note, we do not explore the impact of variation in length of maternity on supervisor-rated job performance outcomes. Data on supervisor evaluations for enlisted Marines are only available for those who remained in service as of October 2017. Therefore we lack sufficient observations of supervisor ratings to study policy changes that took place in 2015 and 2016.

Leave Length & Women's Career Advancement

We explore whether variation in maternity leave length is associated with delays in mothers' promotion trajectories, using the placebo birth matching strategy. We compare mothers to matched nonmothers, broken out by the maternity leave policy at the time of birth or

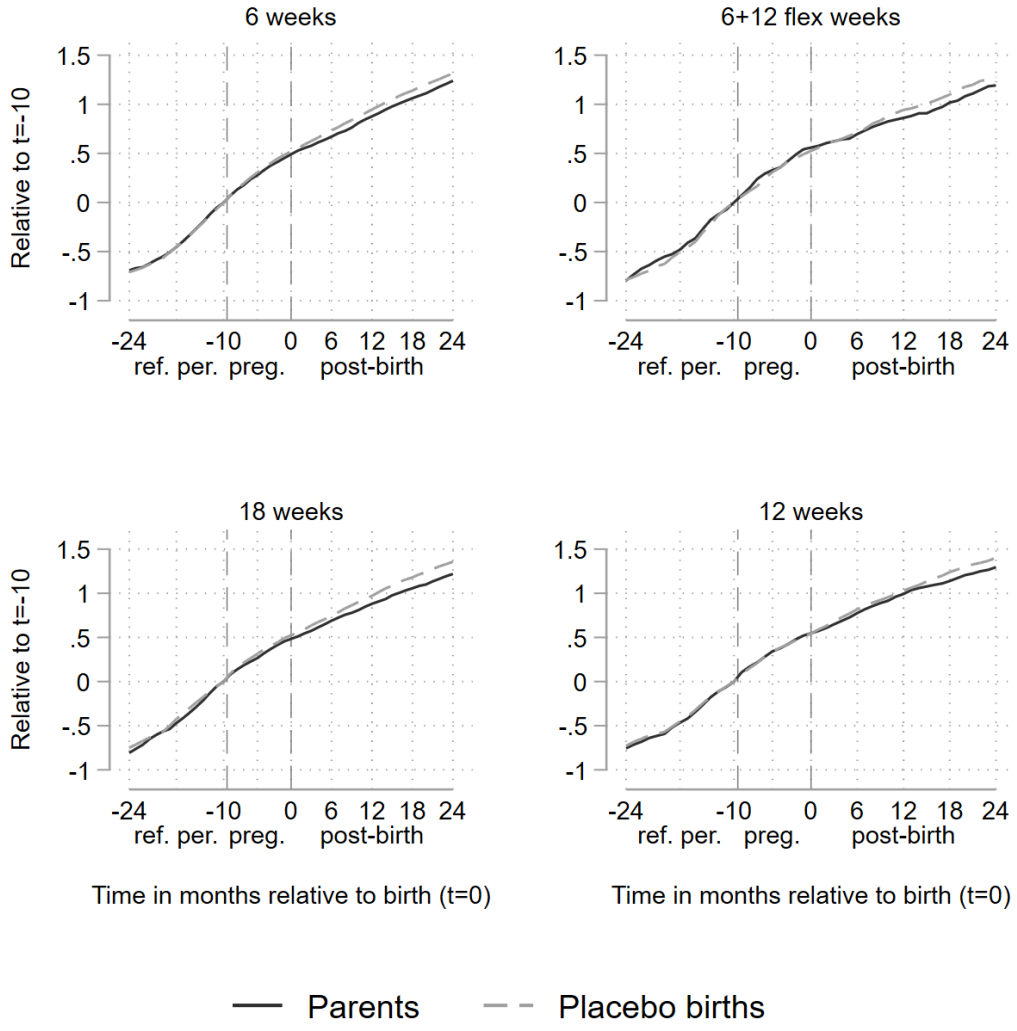
placebo birth. Table 2.3, Panel B displays estimates, and figure 2.5 displays the results graphically.

When measured both 1 month and 1 year after having a child, the largest and only statistically significant declines in number of promotions accrue to women with the *shortest* leave: 6 weeks of time off.[15] Estimates for women with longer leaves at 1 and 12 months postbirth are generally negative in sign but do not reach statistical significance. One exception is among women who received 12 weeks of leave. For these women, the sign of the estimate flips but is not statistically significant. Note, the number of women who gave birth during policy periods with more than 6 weeks of leave is small in size and perhaps not sufficient to detect small impacts. F-tests at 1 month and 12 months postbirth show the size of the group differences across all policy periods is statistically distinguishable from zero, suggesting that leave length matters for promotion, though exactly how is unclear from this analysis.

Two years after giving birth, the direction of the estimated effects suggests all mothers—regardless of time off for leave—are less likely to be promoted. However, estimated impacts are only statistically different from zero for women with 6 weeks and 18 weeks of leave. Again, an F-test of the size of the estimates across all four policy periods differs from zero. Based on the magnitude of the coefficients, it appears that the 12-week leave policy resulted in the smallest declines in number of promotions (roughly 0.03 fewer promotions relative to 0.1 fewer promotions for women who received leaves of 6, 6+12, or 18 weeks of leave). However, results should be interpreted with caution, given that most estimates are not statistically significant.

[15]Note, women can be promoted while on maternity leave. Changes in promotion 1 month postbirth likely reflect performance differences during mothers' pregnancies.

Figure 2.5: Placebo Birth DiD Estimates of Women's Promotion Trajectories by Maternity Leave Length



Notes: Displays promotion trajectories among women, split into subgroups based on the length of paid maternity leave a mother could receive based on the date she gave birth. Policy groups include (a) the initial 6-week leave allowance; (b) the retroactive 18-week policy, where parents expected 6 weeks of leave but actually received an additional 12 weeks of leave (that could be used discontinuously) once back at work; (c) the 18-week policy where mothers had the option to take 18-weeks of leave continuously and largely did (Bacolod et al., 2020); and (d) the policy that afforded 12 weeks of continuous paid leave. See notes on Figure 2.4 for additional details.

2.7 Summary and Conclusions

Using repeated, direct measures of health-dependent work performance combined with longitudinal data on career advancement for service members in the U.S. Marine Corps, we explore the link between the transition to parenthood and workers' outcomes. Our empirical strategy draws on an event study approach based on the precise month of birth as well as a matching design that assigns placebo births to observably similar nonparents. We find both men and women's health-related job performance responds negatively to the transition to parenthood. However, gender differences emerge in the magnitude and persistence of decline and recovery postbirth. Women experience large declines in job-related physical fitness that remain for at least 2 years, while men experience short-lived declines in physical fitness that fade by their child's first birthday. Women's supervisor-rated job performance also declines progressively in the years after having a child, while men's does not. Documented changes in job performance concentrated among women are consistent with our findings that women's promotion trajectories slow while men's do not. Among women, promotion delays accumulate over time; the gap in number of promotions between mothers and nonmothers is largest 24 months postbirth. Promotion trajectories between fathers and nonfathers look almost identical over the 2 years after the birth or placebo birth event.

Last, and perhaps surprisingly, we show that longer paid maternity leave predicts persistent declines in women's health-related physical performance. These findings suggest longer periods away from work may erode job-specific skills. However, in terms of career advancement, we find only suggestive evidence that maternity leave length shapes the degree to which women's promotion trajectories slow after birth. Promotion gaps between mothers and nonmothers are largest and most consistent for women who received the shortest amount of leave—the same group who recover most quickly in terms of physical performance out-

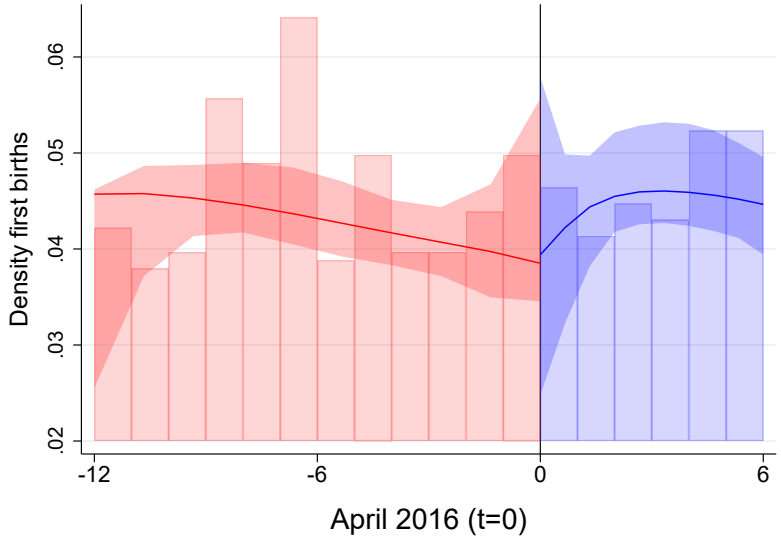
comes. The contrast suggests promotion delays may result from other factors and not from declines in job-related physical performance.

Our findings provide a new angle on the longstanding literature in economics that shows parenthood reduces mothers' employment, hours worked, and wages, while having no effect on fathers. Having a child impacts parents' job performance in the first 2 years of the child's life, highlighting the period after birth as a possible critical window that could give rise to long-term child penalties. Delays in promotion that accumulate for women, but not for men, in the years following birth also underscore the need for increased policy- and firm-level support for recent parents. In our setting, additional parental leave is correlated with larger gender disparities in physical fitness performance. However, longer leave does not translate into larger gender gaps in career advancement across the transition to parenthood. Future research could explore whether alternate family support policies, such as increased access to affordable child care, further mitigate challenges parents face as they transition back into work after having a child.

Appendix

2.A Supplemental Figures and Tables

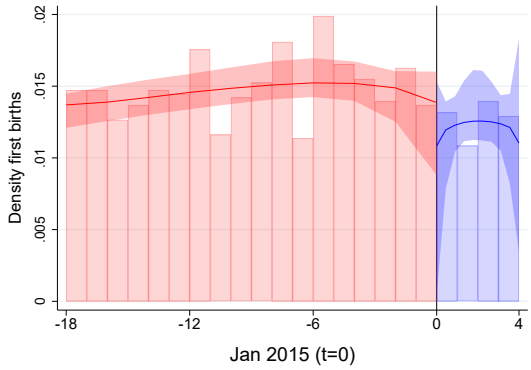
Figure 2.A.1: Density of First Births Under the 18-Week Leave Policy, among Women Pregnant or not yet Pregnant at Time of Announcement



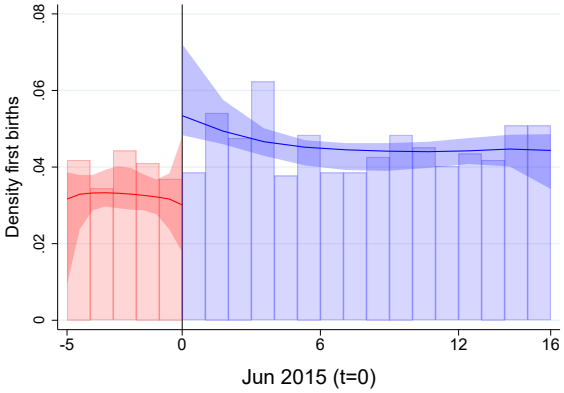
Notes: Histogram bars display the density of first births by month before and after April 2016 (t=0). Births April 2016 and after occur to women whose pregnancies we estimate began *after* the 18-week leave policy announcement. Plotted curves and corresponding 95% confidence intervals come from a manipulation test using a local-polynomial density estimator developed by Cattaneo et al. (2018). The test for a discontinuity at t=0 is not statistically significant. The sample includes all women in the Marines with a first birth during the time window.

Figure 2.A.2: Density of First Births Across Policy Periods

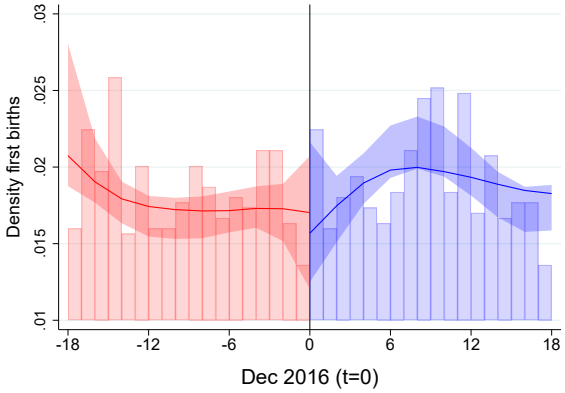
(a) 6 vs. 6 + 12 flex weeks



(b) 6 + 12 flex vs. 18 weeks



(c) 18 vs. 12 weeks



Notes: Histogram bars display the density of first births by month before and after t=0, which differentiates births subject to one leave-length policy period from another. Plotted curves and corresponding 95% confidence intervals come from a manipulation test using a local-polynomial density estimator developed by Cattaneo et al. (2018). The test for a discontinuity at t=0 is not statistically significant in Panel (a), (b), or (c). The sample includes all women in the Marines with a first birth during the time window.

Table 2.A.1: Sample Characteristics of First-Time Parents Across Specifications

	Mothers			Fathers		
First births, obs. required	–	-12/+12mo	-12/+24mo	–	-12/+12mo	-12/+24mo
Age	23.08	23.08	23.46	24.72	25.20	25.49
Education						
Some college	0.05	0.05	0.05	0.05	0.05	0.05
College	0.07	0.08	0.10	0.13	0.15	0.17
Marital Status						
Married	0.67	0.68	0.68	0.85	0.87	0.87
Race/Ethnicity						
African American	0.14	0.15	0.15	0.09	0.09	0.09
Hispanic	0.22	0.22	0.22	0.15	0.15	0.14
Job Classification						
Mngmt./Business/Science/Arts	0.11	0.12	0.14	0.09	0.10	0.10
Service	0.07	0.07	0.07	0.04	0.04	0.04
Sales/Office	0.36	0.36	0.35	0.12	0.12	0.12
Construction/Maint.	0.19	0.18	0.18	0.30	0.29	0.29
Production/Moving/Transpo.	0.20	0.20	0.19	0.13	0.14	0.14
Military	0.07	0.07	0.08	0.32	0.31	0.31
Military Specific Chars.						
Officer	0.05	0.06	0.08	0.11	0.13	0.15
AFQT score (percentile)	57.75	57.87	58.68	61.85	62.61	63.21
GCT score (mean=100; sd=20)	102.69	102.86	103.63	110.25	110.84	111.38
Observations	6747	4502	2801	54479	36212	26916

Notes: Displays characteristics of first-time parents across samples. The columns labeled "–" include all first-time parents. The columns "-12/+12mo" restricts the sample of first-time parents to those observed at least 12 months before and 12 months after the birth, while the columns labeled "-12/+24mo" restrict to those observed at least 12 months before and 24 months after the birth. Time-varying characteristics (e.g., age) are measured 10 months before the birth ($t = -10$). Job categories correspond to Standard Occupational Classification (SOC) system groups applied to U.S. Marine Corps job codes. All averages reflect the share of the sample with the given characteristics, except average age and AFQT and GCT scores. AFQT and GCT scores are measures of intelligence and aptitude, with scoring scales described.

Table 2.A.2: Characteristics of Mothers with Births During the 18-week Leave Policy

	At Time of Announcement		
	Already Pregnant	Not Yet Pregnant	T-stat (test of diff.)
Age	23.66	24.25	-1.438
Education			
Some college	0.06	0.07	-0.589
College	0.12	0.15	-0.864
Marital Status			
Married	0.73	0.69	1.177
Race/Ethnicity			
African American	0.13	0.12	0.522
Hispanic	0.23	0.29	-1.365
Job Classification			
Mngmt./Business/Science/Arts	0.13	0.17	-1.392
Service	0.08	0.09	-0.562
Sales/Office	0.32	0.36	-0.978
Construction/Maint.	0.21	0.20	0.306
Production/Moving/Transpo.	0.18	0.13	1.441
Military	0.09	0.05	1.685
Military Specific Chars.			
Officer	0.11	0.10	0.612
AFQT score (percentile)	56.90	59.28	-1.526
GCT score (mean=100; sd=20)	103.00	104.33	-1.255
Observations	332	205	537

Notes: Displays characteristics of first-time mothers who gave birth during the 18-week leave policy, split by whether or not the woman was already pregnant when the new policy was announced. Per Jukic et al. (2013) pregnancy start estimates rely on an average 268-day gestation. Women already pregnant at time of announcement are those that gave birth within 268 days after the policy announcement. Time-varying characteristics (e.g., age) are measured 10 months before the birth ($t = -10$). Job categories correspond to Standard Occupational Classification (SOC) system groups applied to U.S. Marine Corps job codes. All averages reflect the share of the sample with the given characteristic, except for average age and AFQT and GCT scores, which are raw averages. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 2.A.3: Impact of Childbirth on Physical Performance Test Components

	Physical Fitness Test			Combat Fitness Test		
	(1) 3-Mile Run (time in secs)	(2) Crunches	(3) Pull-Ups	(4) 880-Yard-Run (time in secs)	(5) Lifts	(6) Shuttle Run (time in secs)
			A. Women			
Pregnancy trend	–	–	–	–	–	–
Postbirth drop ($birth - 24mos.$)	43.368***	-3.726***	0.285	8.393***	-1.930*	8.214***
	(5.712)	(0.655)	(1.368)	(1.052)	(0.772)	(1.223)
Recovery trend ($birth - 24mos.$)	-7.945***	0.414	0.381	-1.108***	0.475	-1.309***
	(1.871)	(0.218)	(0.408)	(0.333)	(0.266)	(0.393)
Δ Recovery trend ($13 - 24mos.$)	7.856***	-0.424	-0.582	0.903*	-0.361	1.100*
	(2.287)	(0.265)	(0.503)	(0.404)	(0.286)	(0.485)
12-month effect	11.587**	-2.069***	1.810*	3.962***	-0.029	2.977**
p-value	0.005	0.000	0.029	0.000	0.961	0.001
24-month effect	10.520*	-2.181***	-0.600	1.499	1.341	0.470
p-value	0.028	0.000	0.599	0.083	0.098	0.671
Prepregnancy mean	1520.168	94.233	20.483	215.152	60.638	196.777
Observations	82050	82428	24901	76368	76373	76365
N of individuals	19052	19129	8195	18446	18447	18445
R-squared	0.743	0.576	0.834	0.692	0.537	0.636
			B. Men			
Pregnancy trend	1.654***	0.025**	-0.013***	0.052*	0.002	0.066*
	(0.146)	(0.009)	(0.004)	(0.023)	(0.014)	(0.026)
Postbirth drop ($birth - 24mos.$)	19.571***	0.257***	-0.154***	1.272***	0.400	1.512***
	(1.155)	(0.069)	(0.032)	(0.172)	(0.205)	(0.198)
Recovery trend ($birth - 24mos.$)	-2.263***	0.042***	0.025***	-0.161***	0.038	-0.145***

(Continued on next page)

Table 2.A.3: Impact of Childbirth on Physical Performance Test Components *(Continued)*

	Physical Fitness Test			Combat Fitness Test		
	(1) 3-Mile Run (time in secs)	(2) Crunches	(3) Pull-Ups	(4) 800-Yard-Run (time in secs)	(5) Lifts	(6) Shuttle Run (time in secs)
	(0.136)	(0.009)	(0.004)	(0.020)	(0.021)	(0.023)
Δ Recovery trend $(13-24 mos.)$	2.196***	-0.055***	-0.042***	0.160***	0.009	0.178***
	(0.230)	(0.015)	(0.006)	(0.034)	(0.032)	(0.040)
12-month effect	-5.320***	0.720***	0.117***	-0.504**	0.820***	-0.088
p-value 0	0.000	0.000	0.000	0.001	0.000	0.609
24-month effect	-6.119***	0.567***	-0.086*	-0.526***	1.389***	0.306
p-value	0.000	0.000	0.015	0.005	0.000	0.152
Prepregnancy mean	1324.543	98.870	17.036	174.884	96.249	144.564
Observations	1099159	1103644	1099638	1015207	1015290	1015190
N of individuals	252102	252848	252691	242703	242709	242700
R-squared	0.721	0.634	0.745	0.654	0.578	0.642

Notes: Displays coefficients from the semi-dynamic specification in Eq. 2.2 and the average effect at 12 months and 24 months, with p-values below by item by fitness test type. Physical Fitness Test performance is assessed from January to June. Combat Fitness Test performance is assessed from July to December. Run times are measured in seconds, while crunches, pull-ups, and lifts are measured as raw counts. Physical Fitness Test outcomes in columns (1) to (3) are timed assessments of the activity described. Prior to January 2017, women could opt for an alternative upper-body strength assessment to pull-ups. Limited pull-up outcome data exist for women prior to 2017. Column (4) captures scores on an 880-yard-run, the Movement to Contact drill, designed to mimic the stresses of running under pressure in battle. Column (5) measures the number of times a Marine can lift a 30-pound ammunition can overhead. Column (6) displays timed performance on a 300-yard shuttle run obstacle, Maneuver Under Fire, which includes crawls, ammunition resupply, grenade throwing, agility running, and the dragging and carrying of another Marine. Robust standard errors clustered by ID in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 2.A.4: Impact of Childbirth Across Sample Specifications: Women

First births, obs. required	–	-12/+12mo	-12/+12mo	-12/+24mo	-12/+24mo
No births, restricted to	–	–	⩾24mo in svc	–	⩾36mo in svc
	(1)	(2)	(3)	(4)	(5)
A. Physical Performance Scores					
Pregnancy trend	–	–	–	–	–
Postbirth drop ($birth - 24mos.$)	-0.358***	-0.366***	-0.367***	-0.360***	-0.362***
	(0.025)	(0.027)	(0.027)	(0.031)	(0.031)
Recovery trend ($birth - 24mos.$)	0.046***	0.050***	0.050***	0.055***	0.055***
	(0.008)	(0.009)	(0.009)	(0.010)	(0.010)
Δ Recovery trend ($13 - 24mos.$)	-0.042***	-0.047***	-0.047***	-0.053***	-0.053***
	(0.010)	(0.010)	(0.010)	(0.012)	(0.012)
12-month effect	-0.175***	-0.167***	-0.168***	-0.139***	-0.142***
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
24-month effect	-0.135***	-0.130***	-0.131***	-0.114***	-0.116***
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Prepregnancy mean	-0.002	0.032	0.032	0.140	0.140
N of Individuals	30,883	29,257	25,492	27,728	20,470
Observations	188,677	184,449	174,384	179,261	154,070
R^2	0.601	0.599	0.594	0.597	0.587
B. Supervisor Job Performance Ratings					
Pregnancy trend	0.010**	0.014***	0.013***	0.007	0.006
	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)
Postbirth drop ($birth - 24mos.$)	0.081**	0.119***	0.113***	0.064	0.044
	(0.029)	(0.030)	(0.030)	(0.041)	(0.041)
Recovery trend ($birth - 24mos.$)	-0.027***	-0.031***	-0.032***	-0.017*	-0.018*
	(0.006)	(0.006)	(0.006)	(0.008)	(0.008)
Δ Recovery trend ($13 - 24mos.$)	0.023**	0.029**	0.029**	0.011	0.014
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)
12-month effect	-0.078*	-0.070*	-0.078*	-0.040	-0.064
p-value	[0.014]	[0.033]	[0.017]	[0.305]	[0.099]
24-month effect	-0.119**	-0.104*	-0.110**	-0.113*	-0.114**
p-value	[0.003]	[0.012]	[0.007]	[0.011]	[0.009]
Prepregnancy mean	-0.060	-0.084	-0.084	-0.015	-0.015
N of Individuals	17,105	16,306	13,226	15,547	9,854
Observations	112,958	110,155	101,595	106,171	83,385
R^2	0.438	0.435	0.423	0.433	0.412

Notes: Displays coefficients from the semi-dynamic specification in Eq. 2.2 and the average effect at 12 months and 24 months for various sample specifications. The descriptor "-/+12" restricts to Marines in the sample 12 months before and 12 months after first birth, and "-12/+24" restricts to Marines in the sample 12 months before and 24 months after first birth. "⩾ 24 m.o.s." and "⩾ 36 m.o.s." restricts to Marines who do not experience a birth and have at least 24 or 36 months of service (m.o.s.) respectively. Robust standard errors clustered by ID in parentheses. P-values shown in brackets. $^{***}\ p < 0.001$, $^{**}\ p < 0.01$, $^{*} p < 0.05$.

Table 2.A.5: Impact of Childbirth Across Sample Specifications: Men

First births, obs. required	–	-12/+12mo	-12/+12mo	-12/+24mo	-12/+24mo
No births, restricted to	–	–	⩾24 in svc	–	⩾36 in svc
	(1)	(2)	(3)	(4)	(5)
A. Physical Performance Scores					
Pregnancy trend	-0.009***	-0.006***	-0.006***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Postbirth drop ($birth - 24mos.$)	-0.138***	-0.090***	-0.092***	-0.065***	-0.067***
	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
Recovery trend ($birth - 24mos.$)	0.009***	0.006***	0.006***	0.009***	0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Δ Recovery trend ($13 - 24mos.$)	-0.009***	-0.006***	-0.006***	-0.011***	-0.011***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
12-month effect	-0.041***	-0.027***	-0.027***	0.031***	0.027***
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
24-month effect	-0.038***	-0.028***	-0.028***	0.002	0.001
p-value	[0.000]	[0.000]	[0.000]	[0.746]	[0.857]
Prepregnancy mean	0.202	0.245	0.245	0.295	0.295
N of Individuals	366,120	349,290	311,878	340,874	270,630
Observations	2,438,278	2,366,516	2,265,486	2,316,730	2,066,435
R^2	0.602	0.600	0.596	0.600	0.594
B. Supervisor Job Performance Ratings					
Pregnancy trend	-0.001	0.004**	0.003**	0.004**	0.004*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Postbirth drop ($birth - 24mos.$)	-0.028**	0.018	0.016	0.028*	0.020
	(0.009)	(0.010)	(0.010)	(0.011)	(0.011)
Recovery trend ($birth - 24mos.$)	0.000	-0.002*	-0.002*	0.002	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Δ Recovery trend ($13 - 24mos.$)	0.000	0.001	0.001	-0.006**	-0.005*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
12-month effect	-0.025**	-0.008	-0.008	0.051***	0.042***
p-value	[0.007]	[0.450]	[0.431]	[0.000]	[0.000]
24-month effect	-0.020	-0.023	-0.021	0.003	0.009
p-value	[0.080]	[0.070]	[0.091]	[0.828]	[0.466]
Prepregnancy mean	0.034	0.016	0.016	0.011	0.011
N of Individuals	178,679	171,825	143,426	167,732	112,968
Observations	1,265,844	1,238,183	1,152,795	1,216,340	986,863
R^2	0.434	0.431	0.427	0.430	0.422

Notes: Displays coefficients from the semi-dynamic specification in Eq. 2.2 and the average effect at 12 months and 24 months for various sample specifications. The descriptor "-/+12" restricts to Marines in the sample 12 months before and 12 months after first birth, and "-12/+24" restricts to Marines in the sample 12 months before and 24 months after first birth. "⩾ 24 m.o.s." and "⩾ 36 m.o.s." restricts to Marines who do not experience a birth and have at least 24 or 36 months of service (m.o.s.) respectively. Robust standard errors clustered by ID in parentheses. P-values shown in brackets. $^{***}\ p < 0.001$, $^{**}\ p < 0.01$, $^{*} p < 0.05$.

Table 2.A.6: Sample Characteristics of First-Time Mothers Across Policy Periods

	Paid Leave Length				
	6 weeks	6 +12 flex weeks	18 weeks	12 weeks	p-value (F-test of diff.)
Age	23.31	23.76	23.89	23.35	0.06
Education					
Some College	0.05	0.05	0.07	0.05	0.44
College	0.09	0.12	0.13	0.10	0.14
Marital Status					
Married	0.67	0.70	0.72	0.67	0.33
Race/Ethnicity					
Black	0.16	0.12	0.13	0.18	0.18
Hispanic	0.19	0.21	0.26	0.30	0.00
Job Classification					
Mngmt./Business/Science/Arts	0.13	0.16	0.14	0.15	0.45
Service	0.06	0.07	0.08	0.07	0.42
Sales/Office	0.37	0.28	0.33	0.32	0.06
Construction/Maint.	0.17	0.20	0.20	0.17	0.52
Production/Moving/Transpo.	0.19	0.21	0.16	0.22	0.18
Military	0.08	0.08	0.08	0.06	0.79
Military Specific Chars.					
Officer	0.07	0.09	0.11	0.08	0.12
AFQT score (percentile)	59.31	60.13	57.81	56.70	0.04
GCT score (mean=100; sd=20)	103.85	104.70	103.51	102.48	0.19
N	1653	135	537	415	2801

Notes: Displays characteristics of first-time mothers in our preferred sample by maternity leave length. Policy groups include (a) the initial 6-week leave allowance; (b) the retroactive 18-week policy, where parents expected 6 weeks of leave but actually received an additional 12 weeks of leave (that could be used discontinuously) once back at work; (c) the 18-week policy where mothers had the option to take 18-weeks of leave continuously and largely did (Bacolod et al., 2020); (d) and the policy that afforded 12 weeks of continuous paid leave. The column labeled p-value (F-test of diff.) displays the p-value of an F-test for differences across all policy groups. See Table 1 for notes on variables included.

Chapter 3

Variation in Teacher Quality over the Preschool Year and its Implications for Early Childhood Education Accountability Systems

3.1 Introduction

Research consistently shows early childhood education (ECE) programs improve children's short- and long-term outcomes. Yet, not all ECE programs are created equal. Scholars increasingly highlight the need for *high-quality* programs in order for children to reap the benefits of ECE (Dodge, 2017). In turn, policymakers have put in place accountability systems to monitor and increase the overall effectiveness of ECE programs, as well as narrow disparities in program quality. Unlike K-12 education accountability systems, which monitor

students' academic outcomes, ECE accountability systems assess whether programs have key inputs in place to promote children's development. Measured inputs tend to include static features of classrooms, such as teacher qualifications and class size, as well as in-person evaluations of the dynamic day-to-day classroom environment.

The success of an inputs-based approach to accountability rests on monitoring the right inputs—those that effectively promote young children's development—and on measuring them accurately. One popular indicator for ECE accountability purposes is the Classroom Assessment Scoring System (CLASS^{TM}) (Pianta et al., 2008). The CLASS is an observational teacher assessment tool that quantifies the extent to which teacher-student interactions promote children's development. Most state-level accountability systems, called Quality Rating and Improvement Systems (QRIS), rely at least in part on the CLASS to benchmark ECE provider quality across the state (National Center on Early Childhood Quality Assurance, 2017). At the federal level, the Head Start Designation Renewal System also uses average teacher CLASS scores to evaluate the quality of more than 1,500 publicly funded Head Start agencies—those that provide free early education services to families below the poverty line.

ECE accountability systems, including QRISs and the Head Start Designation Renewal System, tie consequences and rewards to measured program performance. For example, QRIS make programs' quality ratings public to put pressure on poor performing providers to improve the quality of services or risk going out of business. QRIS ratings also identify programs in need of technical assistance or those eligible for financial bonuses. The Head Start Designation Renewal System imposes high-stakes funding consequences on programs that do not meet quality standards. Head Start agencies deemed poor performing lose their grant funding and have to recompete with other agencies to be reinstated.

Quality measurement in ECE accountability systems plays an important role in determin-

ing which programs are penalized or rewarded. When the CLASS is used as an accountability benchmark, on-site ratings typically occur during a single program visit over a few days or a single week. The assumption behind this "point-in-time" measurement approach is that teacher quality is stable throughout the year: whether assessed in the fall, winter, or spring, teachers in a program would generally receive the same scores. Yet, limited empirical research tests this assumption that teaching quality is stable. If the quality of instruction in classrooms rises or falls within the school year, then linking point-in-time CLASS scores to positive or negative accountability incentives may incorrectly reward or penalize programs, depending on when each program is observed. Understanding the extent to which "point-in-time" estimates accurately capture teacher quality over the course of the year is critical for ensuring the fairness and accuracy of ECE accountability systems.

The present study offers insight into whether point-in-time CLASS observations adequately categorize ECE program quality for accountability purposes. Our analysis draws on rich data with multiple observations of teachers' CLASS scores over a single preschool year to explore whether and how observed classroom quality varies. We use flexible multilevel growth curve models to estimate teacher-level trajectories of change in CLASS scores across three measured domains of classroom quality. We quantify our findings in two ways. First, we decompose variability in a single teacher's CLASS scores over the year into variation based on predicted time trends vs. other time-invariant sources of variation. Second, we explore the implications of CLASS score time trends for ECE accountability systems, using the federal-level Head Start accountability policy as a test case. Specifically, we build a simulation model to estimate a Head Start agency's likelihood of scoring below one of three given CLASS score thresholds, based on their assessment timing during the year and our growth curve findings.

Results indicate that teachers' CLASS scores vary systematically over the course of a school year, particularly in instructional support (i.e., engaging students in higher-order thinking). In our sample, quality ratings of instructional support practices decline during the fall (August–December) and improve during the winter (January–March). These time-patterned trends in instructional support explain nearly one quarter (24%) of what would otherwise be characterized as random or unpredictable teacher-level fluctuation in instructional quality. Additionally, simulation analyses show that seasonal patterns in instructional support CLASS scores have the potential to affect accountability outcomes for Head Start programs at the lower end of the quality distribution. For a lower-performing Head Start program, their likelihood of falling below the instructional support quality threshold varies sizably depending on whether they are evaluated in the fall, winter, or spring.

Notably, the consequences of misclassification under ECE accountability systems are meaningful. For example, in Head Start, poor performance on any one CLASS domain requires Head Start agencies to compete with outside applicants for renewed funding. Competition is both time and resource intensive and requires Head Start programs to juggle ongoing service provision while drafting their grant application.

3.2 Observational Measures in ECE Accountability Systems

ECE accountability systems rely on two types of measures to benchmark quality: program/classroom features that either (1) indirectly or (2) directly promote sensitive, responsive, and cognitively stimulating interactions between teachers and children. Indirect measures of ECE quality, known as structural quality indicators, include aspects of teachers,

classrooms, and programs that set the stage for effective teaching and learning. Structural quality indicators are generally easy-to-measure, static features of the ECE environment. Examples include teachers' educational qualifications, the size of classes, the ratio of number of staff to number of children, or the types of educational materials used in classrooms. Direct measures of ECE quality, termed process quality measures, include observational assessments of teacher–child interactions or peer-to-peer interactions.

ECE accountability systems commonly rely on the CLASS to assess the quality of teacher-child interactions and benchmark process quality. The CLASS aligns closely with developmental theory on children's early skill acquisition (Hamre et al., 2008). It quantifies the nature of students' day-to-day experiences in the classroom through observations and subsequent ratings of teacher–child interactions and classroom processes. In the current paper, we refer to the CLASS as a measure of teacher quality, given that teachers largely drive the classroom operations, interactions, and climate that give rise to classroom-level quality. However, we acknowledge that classroom quality is multifaceted and that teachers do not drive all the classroom-level factors that determine CLASS scores.

CLASS ratings are organized into three domains: emotional support, classroom organization, and instructional support. Each of the domains includes smaller dimensions of quality, under which specific behavioral indicators are defined. Broadly, emotional support ratings measure how well a classroom environment and a teacher's practices respond to students' needs, emotions, and interests. The classroom organization domain quantifies how well teachers manage the classroom so that students can self-regulate and focus their behavior and attention on academic goals. The instructional support domain evaluates whether teaching practices foster students' concept development, analytic reasoning skills, and abstract thinking. Strong instructional support practices include asking open-ended questions, modeling

rich language, and providing specific feedback to students to expand understanding.

Teacher ratings on the CLASS consistently predict improvements in children's outcomes, though these findings come largely from correlational studies and show only moderate to small associations (Araujo et al., 2016, Perlman et al., 2016). Across CLASS domains, emotional support (i.e., responsive teaching) is most closely linked to children's social–emotional outcomes; classroom organization (i.e., positive management and routines) is most closely linked to students' inhibitory control and self-regulation; and instructional support (i.e., cognitive facilitation) is most closely linked to early language and literacy development (Downer et al., 2010, Hamre et al., 2008, 2014). On average, preschool teachers receive high emotional support and classroom organization ratings and lower instructional support ratings (Hamre et al., 2008, Office of Head Start, 2013, Perlman et al., 2016).

Domain-specific CLASS ratings feature prominently in ECE accountability systems. For example, 44% of state-based QRISs assess teachers on the CLASS to determine an ECE program's overall quality rating. In some states, average program-level CLASS scores for each domain, in combination with performance on other specified indicators, help place programs into different quality tiers (e.g., a 5-star program must be above a given threshold on each CLASS domain) under the QRIS. Other states use average domain-specific CLASS scores to award quality points to a program that then count toward the program's overall QRIS rating (e.g., a certain number of quality points awarded if a program scores above a given threshold on each CLASS domain). Under Head Start accountability policy, average teacher CLASS scores in each quality domain, in addition to six other performance standards, determine whether a Head Start agency "passes" or "fails" its accountability review.[1] The

[1]The terms "pass" and "fail" are not used by the Office of Head Start nor the Head Start Designation Renewal System policy. We use these terms in the present study for simplicity to differentiate between when a Head Start agency performs well enough to continue to receive funding from when it underperforms and must reapply for funding to continue to operate.

Head Start Designation Renewal System designates programs that fail to meet any one of these seven performance standards as poor performing. Poor performance on the CLASS was the second most common reason Head Start agencies failed their accountability reviews during the rollout of the initial policy between 2011 and 2014 (Office of Head Start, 2016). In that period, 140 Head Start agencies were required to compete for renewed funding due to CLASS scores below required thresholds.

State QRISs and Head Start Designation Renewal System evaluations have important consequences for ECE programs. Under QRISs, ECE programs' quality ratings are made public, and recent research shows low QRIS ratings can reduce enrollment and drive poorly rated programs out of business (Bassok et al., 2019). High performance ratings under QRISs often qualify programs for financial rewards, such as program-level bonuses or higher reimbursement rates under the child care subsidy system. Within Head Start, if a program fails its accountability review, it must reapply and compete against outside applicants for continued grant funding. The reapplication process takes roughly 18 months, at which time the incumbent Head Start agency either wins back its grant funding or a new agency is awarded the grant. During the reapplication process, the incumbent agency continues to provide Head Start services without additional funds to support the reapplication process. In contrast, Head Start agencies that pass their accountability reviews are granted automatic funding renewal and face no additional administrative burdens.

While classroom observation assessment tools, such as the CLASS, hold promise as measures of ECE process quality, their effective use in high-stakes accountability systems relies on several important assumptions. For example, current approaches to ECE accountability assume that CLASS ratings among a subset of teachers in a program effectively capture variation in *program-level* effectiveness when averaged together (Sabol et al., 2019). Sim-

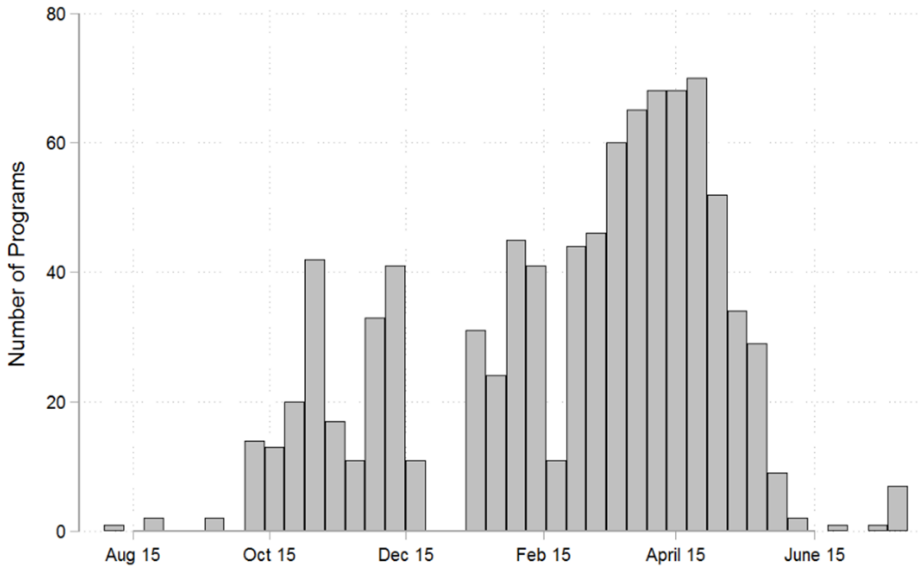
ilarly, QRISs and the Head Start Designation Renewal System generally conduct CLASS evaluations for ECE programs every three years, assuming that year-to-year variation is not of concern. Mashburn (2017) outlines several other assumptions baked into the logic of the Head Start Designation Renewal System, highlighting that weak empirical evidence for many of the underlying assumptions may compromise the accuracy and fairness of resulting accountability decisions.

However, one key assumption not previously considered involves *within*-school-year variation in teaching quality. For practical reasons, under QRISs and the Head Start Designation Renewal system, programs receive their on-site accountability reviews at different times during the school year. Figure 3.1 shows the variation in timing of on-site visits and ensuing CLASS assessments across different Head Start agencies during the initial rollout of the Head Start Designation Renewal System, between 2012 and 2014. If teaching quality in ECE classrooms evolves during the school year, then accountability systems like the Designation Renewal System may arbitrarily rate some programs higher or lower than others due to variation in the time of year programs are assessed.

3.3 Why Would Preschool Teacher Quality Vary Within the School Year?

There are several theoretical reasons to expect preschool teacher quality to vary over the school year. For example, as the year progresses, teachers and students may adjust to one another and develop rapport, increasing the effectiveness of educational interactions. Similarly, novice teachers may need time to master effective practices (Malmberg et al., 2010) while experienced teachers may need time to tailor practices to each new group of students

Figure 3.1: Head Start Agency CLASS Observation Dates Under the Head Start Designation Renewal System (2012–2014)



Notes: Displays the number of Head Start grantee agencies with on-site observations by time over the school year under the Head Start Designation Renewal System rollout between 2012 and 2014. Dates of the on-site review are coded as the first day federal reviewers began the visit. Data come from public records available through the Program Locator feature on the Head Start Early Childhood Learning and Knowledge Center website.

at the start of the year. On the other hand, with time, teachers may experience burn out, undermining the quality of their interactions with students as the year progresses. Finally, external factors linked tightly to calendar time, such as a holidays or seasonal patterns of attendance (e.g., increased absences in the winter), may shape the composition of students in a classroom, the content taught, or a teacher's own capacity to engage with students (Malmberg et al., 2014). In line with these theories, some evidence suggests student learning gains across elementary and middle school grades are nonlinear, decelerating as the school year progresses (Kuhfeld and Soland, 2021, von Hippel and Hamrock, 2019).

In addition, theory suggests teachers' instructional support practices may show the greatest change over time, more so than teaching practices around classroom organization or

emotional support. First, strong instructional support practices, such as posing open-ended questions and building on students' prior knowledge, largely depend on the teaching content. As content shifts within a classroom over the school year, a teacher's capacity to provide this type of instructional support is more likely to change (Praetorius et al., 2014). Second, Curby et al. (2011) suggest that high-quality instructional interactions require other key classroom practices already in place. For example, in ECE contexts, children may need to feel emotionally secure in the classroom before they can engage in cognitively demanding tasks. In addition, if teachers learn to manage their classrooms effectively, then routines and activities will run smoothly and instruction will proceed more effectively (e.g., Bohn et al., 2004). In this case, instructional support CLASS ratings may evolve more over time than emotional support or classroom organization ratings and improvements may take longer to accrue.

3.4 Evidence on Variation over Time in Observed Teacher Quality

Studies on variation in teacher quality typically focus on a short length of time (i.e., multiple observations within a day) or long length of time (i.e., multiple years) but with observations spread out over time and thus sparse in number. Few studies include ratings of an adequate number of teaching occasions over a sufficiently wide time frame to identify *trajectories of change* in the quality of teachers' instructional practices over a long time window. The challenge is that very few data sets exist with month-to-month or week-to-week observations of the same teachers for an extended period of time, limiting our ability to study this question well.

Studies that focus on wide windows of time with a low density of teacher ratings (e.g., 6-10 observations across 3 years (Malmberg et al., 2014) suggest variation exists in preschool teacher quality over time. For instance, Castro et al. (2017) and Buell et al. (2017) find preschool teachers' CLASS scores generally increase between the fall and spring when measured at each juncture. However, patterns are not always consistent across year nor CLASS domains (Buell et al., 2017), and with just 2 observations per teacher per year it is difficult to parse out true change in teaching quality from idiosyncratic fluctuations in ratings.

Recognizing this limitation, other studies have taken care to limit analyses to a small observation window with a higher density of rated snapshots of preschool teachers (e.g., four back-to-back CLASS ratings during the first two hours of the preschool day; Curby et al., 2010. However, this approach isolates a period of time where changes in observed quality may reflect changes in instructional activities or content focus over the course of the day (Wang et al., 2020), rather than actual changes in teaching quality.

A handful of studies, similar in nature to ours, rely on high-frequency preschool teacher ratings during a school year to explore within-year differences in teaching quality. Both Cash and Pianta (2014) and Meyer et al. (2011) divide the year into discrete time periods (quarters or months) and document differences in CLASS ratings across these time points. Generally, they find improvements in teacher quality ratings as the school progresses. However, these studies stop short of characterizing time trends over a continuous set of observation dates during the school year, instead of averaging over fixed time periods. In contrast, other studies do trace trajectories of change in observed teacher quality across the school year but focus on middle and high school contexts, finding mixed patterns of results (Briggs and Alzen, 2019, Casabianca et al., 2013, 2015).

The paper most similar to ours explores within-year variation in ECE program-level

CLASS ratings using a cross-sectional approach. Gandhi et al. (2020) use data on CLASS scores from publicly funded pre-Kindergarten programs in New York City. They then compare how preschools rated at different time points during the year vary in terms of teachers' average CLASS scores. They find programs rated in the fall have lower CLASS scores compared to those rated in the spring, controlling for program characteristics (e.g., program type and location) and characteristics of students attending the programs (e.g., race/ethnicity and poverty status). However, programs evaluated in the fall were primarily community-based, whereas those evaluated in the spring were located in public schools. Due to the cross-sectional nature of the data, the authors cannot fully disentangle whether time patterns in ratings are driven by the timing of the CLASS observation itself or unequal distribution regarding when certain types of programs are observed.

3.5 Current Study

Our study builds on and extends the existing literature base in several key ways. First, we leverage repeated measures for the same teachers over time to understand within-teacher trajectories of change in quality ratings. The within-teacher approach minimizes the extent to which differences between teachers (or between programs, as in Gandhi et al. (2020)) may drive results. While prior research uses a within-teacher approach to study time trends in teaching quality among middle and high school teachers, this paper is the first to do so in the ECE context. Second, we measure CLASS observation timing at a more granular level than available in previous ECE focused studies. Precise observation dates combined with data across most days of the year also allow us to flexibly characterize the shape of growth and decline in teacher-level CLASS ratings, avoiding researcher-imposed functional form specifications.

Finally, our study adds a broader understanding of the significance of time-patterned trends in preschool teacher quality to overall quality measurement. Using a variance decomposition, we explore to what extent within-year time trends in CLASS scores explain total variation in teachers' CLASS scores. Put differently, are within-teacher fluctuations in rated performance in part explained by a teacher's growth or decline in teaching quality over the school year? Last, we use our teacher-level findings on time trends to build a simulation model that aggregates teacher-level time trends in quality to program-level growth and decline. We then compare simulated Head Start program-level CLASS scores to accountability thresholds, quantifying whether time trends in teaching quality are sizeable enough to make a difference in programs' quality ratings.

3.6 Data and Methods

Our analytic sample draws on data from the National Center for Research on Early Childhood Education's (NCRECE) Teacher Professional Development Study (Pianta and Burchinal, 2007). Teachers in our sample were recruited as part of the NCRECE Study from large community preschools and Head Start programs across nine socioeconomically diverse U.S. cities. Study-eligible teachers met the following criteria: (1) they were the lead teacher in a classroom where at least half the students would attend kindergarten the following year; (2) they provided the majority of classroom instruction in English; and (3) they had access to internet in their schools to upload videos of classroom instruction to an online platform for study purposes. Teachers were located in: New York City, NY; Hartford, CT; Chicago, IL; Stockton, CA; and Dayton, OH (all recruited in spring 2008); and Columbus, OH; Memphis, TN; Charlotte, NC; Providence, RI; and again, Chicago, IL (all recruited in spring 2009).

The present analysis focuses on data collected during the second phase of the two-phase

NCRECE Study, during which teachers were randomly assigned to either a coaching condition (treatment) or business-as-usual condition (control). During this second phase, teachers recorded and regularly uploaded videos of their classroom instruction. Teachers randomized to the professional development treatment, worked with coaches who watched videos of their classroom instruction, highlighted effective teaching techniques, and discussed strategies to improve the quality of teachers' interactions with students. Teachers in the control group submitted videos of their classroom instruction but maintained teaching practices as usual. Teachers' videos were scored on the CLASS and used to evaluate the effectiveness of the teacher professional development coaching intervention.

We also draw on data from the 2009 Head Start Family and Child Experiences Survey (FACES) to help address our last research question on the importance of within-year time trends in CLASS scores for accountability outcomes. The FACES 2009 provides a nationally representative sample of Head Start children, families, classrooms, and programs. Importantly for our purposes, Head Start teachers in the 2009 survey were rated on the CLASS. Raters assessed all teachers in Spring of the 2009–2010 school year. CLASS ratings from the FACES 2009 are available for multiple Head Start teachers within a single Head Start agency. We use this data to estimate the degree of correlation in teachers' CLASS scores within a Head Start agency for use in our simulation analyses.

3.6.1 Sample Characteristics

Our total possible analytic sample includes 401 preschool teachers who were randomly assigned in phase two of the NCRECE Study to either the professional development intervention or a control condition. We restrict our sample to 303 teachers for whom we observe outcome data (i.e., at least one CLASS score in each rated domain). Characteristics of teach-

ers in our analytic sample as compared to all teachers in the NCRECE Study are presented in Table 3.1. Preschool teachers in our sample have an average of 15 years of experience teaching preschool, are on average 42 years old, and have completed 16 years of education on average (roughly equivalent to a bachelor's degree). More than half of the teachers (57%) report teaching in a preschool program that offers Head Start and another 38% report teaching in a preschool program housed within a public school. (Note, these measures are not mutually exclusive.) Among teachers in the sample, 45% identify as Black, 32% as white, 12% as Hispanic, and 7% as another race. Students in teachers' classrooms are predominantly from low-income families: 88% of students in the average classroom have household income that falls below an income-to-needs ratio of two. We include both treatment and control teachers from the phase two NCRECE intervention in our analytic sample. In our sample, 55% of teachers were assigned to ongoing professional development.

Teachers in our analytic sample were comparable to the full sample of teachers that participated in the second phase of the parent study on most observed characteristics with the following exceptions: those in the analytic sample (1) have slightly more years of education, (2) are more likely to be white and less likely Hispanic, and (3) more likely assigned to professional development coaching.

3.6.2 Measures

Observer-Rated Teacher Quality

Our measure of teaching quality comes from observational ratings of teachers, scored using the CLASS instrument across three measured domains of quality: emotional support (e.g., the positivity of the learning climate), classroom organization (e.g., the structure of lessons, rules, and routines), and instructional support (e.g., engagement of students in higher-order

Table 3.1: Characteristics of Teachers and Classrooms in the Full and Analytic Samples

	Full Sample Mean (sd)	Analytic Sample Mean (sd)	Excluded Sample Mean (sd)	Difference: Excluded Analytic (t-statistic)
Total years of experience teaching	14.160	14.567	12.756	-1.811
	(9.134)	(9.526)	(7.509)	(-1.604)
Teacher age	42.337	42.437	42.000	-0.437
	(10.618)	(10.936)	(9.522)	(-0.337)
Teacher years of education	15.742	15.867	15.322	-0.545**
	(1.651)	(1.632)	(1.653)	(-2.727)
Classroom in a Head Start center	0.574	0.570	0.632	0.061
	(0.495)	(0.496)	(0.496)	(0.520)
Classroom in a public school	0.370	0.382	0.200	-0.182
	(0.484)	(0.487)	(0.410)	(-1.632)
Share of teachers who identify as:				
Black	0.439	0.452	0.398	-0.054
	(0.497)	(0.499)	(0.492)	(-0.938)
White	0.292	0.323	0.194	-0.130*
	(0.455)	(0.469)	(0.397)	(-2.465)
Hispanic	0.140	0.119	0.204	0.085*
	(0.347)	(0.324)	(0.405)	(2.124)
Another race/ethnicity	0.075	0.069	0.092	0.023
	(0.263)	(0.254)	(0.290)	(0.736)
Classroom-level poverty (% of students)	0.880	0.878	0.903	0.025
	(0.210)	(0.213)	(0.169)	(0.577)
Share of teachers assigned to professional development	0.511	0.554	0.378	-0.177**
	(0.500)	(0.498)	(0.487)	(-3.074)
Observations	401	303	98	401

Notes: Columns 1–3 present means for the full, analytic, and excluded samples; standard deviations are shown in parentheses. Column 4 displays the difference in means between the analytic sample and those excluded; t-statistics are shown in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$.

thinking). Scores on each of these three CLASS domains fall on a seven-point scale. Scores of 1– 2 are generally considered low quality, 3–5 mid-range quality, and 6–7 high quality.

CLASS score ratings in our sample come from assessments of in-person and video-recorded teacher instruction. Each teacher observation was split into two 20-minute segments. Each 20-minute segment was then double-coded by two randomly assigned and certified CLASS raters. Scores from each rater were averaged to produce a single average CLASS score per segment. Average segment ratings were then averaged to produce a single average CLASS domain score rating for a given date of instruction.

We observe 2,308 teaching video submissions from the 303 teachers in the analytic sample, resulting in an average of 7.6 observations rated using the CLASS per teacher throughout the school year. Table 3.2 presents means and standard deviations of CLASS scores in the analytic sample, broken out by teachers receiving professional development and those under business-as-usual conditions.

Table 3.2: Descriptive Statistics on Analytic Sample CLASS Scores

					Percent of videos observed by season		
CLASS Domain		Mean	SD	# of videos	Fall (Aug 12-Dec 31)	Winter (Jan 1-Mar 31)	Spring (Apr 1-Jul 1)
Emotional Support	Receiving PD	5.33	0.63	1,624	39%	33%	28%
	Business as usual	5.14	0.58	684	46%	32%	21%
Classroom Organization	Receiving PD	5.38	0.67	1,624	39%	33%	28%
	Business as usual	5.36	0.61	684	46%	32%	21%
Instructional Quality	Receiving PD	2.42	0.82	1,624	39%	33%	28%
	Business as usual	2.10	0.61	684	46%	32%	21%

Notes: Displays the mean and standard deviation of CLASS score by domain across the preschool year among teachers receiving professional development (PD) over the school year, and those teaching under usual conditions. The number of videos that contribute to these average estimates is displayed broken out by group, as well as the share of videos that fall into each season during the school year.

Observation Dates

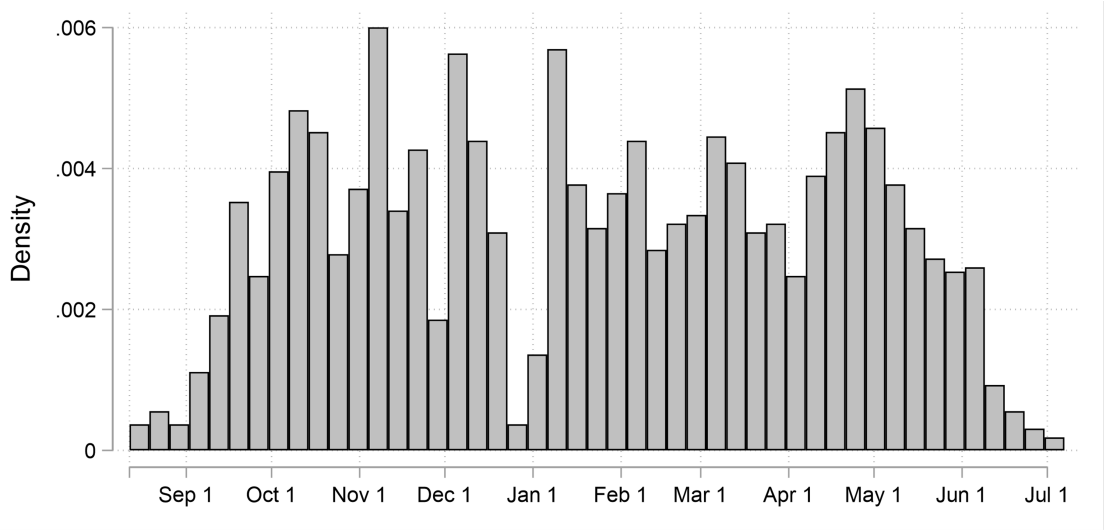
We measure the date of a teacher's in-person or video-recorded observation using a continuous measure of months rather than days for ease of interpretation. Teachers in our sample were observed and rated during the 2008–2009 school year or 2009–2010 school year. The year we observed each teacher's CLASS scores depends on when study recruitment took place in their city and, subsequently, the year in which the phase two intervention took place. We combine both years of data (2008–2009 and 2009–2010) to map within-school-year time trends in CLASS scores. In this way, we treat observations on dates in August 2008 and August 2009 as reflective of general "August" date scores.

Figure 3.1 shows the distribution of CLASS score observations over the preschool year. There is wide coverage across dates over the school year, with the exception of the end of December to early January when preschools are generally not in session, due to Christmas and the New Year holidays.

Covariates

Our analyses account for differences in classroom characteristics across a set of teacher and classroom covariates that might influence a teacher's average CLASS score. We control for each teacher's level of experience, measured as the total number of years teaching preschool. We also control for teacher age; teacher race/ethnicity, coded as Black, white, Hispanic, or other; and teachers' level of education, coded as the number of years of education they have obtained. Binary indicators capture whether the teacher reports teaching in a public school or a Head Start program. We proxy for classroom-level student poverty composition by measuring the share of students in the classroom who fall below an income-to-needs ratio of two. Last, we construct indicators to control for features of the NCRECE Study design.

Figure 3.1: Distribution of CLASS Observation Dates Across Teachers in the Sample Over the Preschool Year



Notes: Displays the density of video-based teacher observation dates in the analytic sample over the school year, based on a 1-week bins.

We include a binary indicator for whether teachers were randomly assigned to a 14-week professional development course in the first NCRECE intervention phase, as well as a binary indicator for whether the teacher was newly added to the NCRECE Study in phase two.

3.6.3 Empirical Strategy

Trends in CLASS Scores over the Preschool Year

Our main approach fits separate multilevel linear regressions, also known as linear splines, to different ranges of the data. We use this modeling approach for two reasons. First, preliminary analyses displayed poor model fit at both the beginning and end of the distribution of calendar time when we used a range of parametric models (e.g., linear, quadratic, and cubic curves). Second, without strong evidence on the functional form for within-year change in teaching quality, we did not want to introduce bias by choosing a linear or polynomial

curve that might be misspecified. Our spline approach ensures each observation affects the relevant local slope parameter rather than the entire fit of the curve across the full range of calendar time. Through this approach, we achieve better model fit and avoid imposing researcher-specified functional form assumptions to characterize trajectories of change in CLASS scores.

We divide time during the preschool year into three distinct segments using two knot points. Knot points are placed at December 31 and March 31. We chose the number and location of the knot points, based on observed patterns in the data, modeled using both locally weighted scatterplot smoothing (lowess) curves and local polynomial and fractional polynomial curves (see Figure 3.A.1). Knot points at the end of December and March also roughly divide the school year into meaningful time periods: fall (August–December), winter (January–March), and spring (April–July).[2]The resulting linear spline model produces three separate seasonal slope parameters that together characterize whether and how a typical teacher's observer-rated teaching quality varies over the school year.

We use two-level, random-slope, and random-intercept hierarchical linear models (HLM) to fit our linear piecewise regressions. The HLM approach accounts for the fact that repeated measures of CLASS scores are nested within teachers/classrooms. It also allows us to estimate teacher-level residual variation around the average slope for all teachers, a teacher-specific random intercept term around the average intercept for all teachers, and a residual term that represents remaining unexplained variation in observed classroom quality for multiple observations within a teacher.

Because some teachers in our sample received individualized professional development

[2]We tested whether our modeling approach is sensitive to the location of these knot points as well as to excluding a small number of classroom observations during the summer months (August, June, and July). The results are not sensitive to either specification check.

coaching as part of the NCRECE Study's phase two intervention, we model different trajectories of change for those assigned to professional development and those assigned to the control group, or "business-as-usual" condition. Our approach thus separately identifies how CLASS scores evolve over the school year when teachers are engaged in intensive professional development from how they evolve over the school year for teachers under business-as-usual teaching conditions. We consider the time trend results among teachers receiving professional development as a possible upper bound on CLASS score growth over the year.

Our general multilevel linear piecewise regression used to model time trends in individual teachers' CLASS scores is presented below.

For teacher j in city c at time point t, we have:

Level 1: Time (t)

$$CLASS_{jtc} = \beta_0 + \beta_{1j}Fall_t + \beta_{2j}Winter_t + \beta_{3j}Spring_t + \vec{X_j}\vec{\beta_4} + a_c + \epsilon_{jtc} \qquad (3.1)$$

Level 2: Teacher (j)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}ProfDev_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}ProfDev_j + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}ProfDev_j + u_{1j}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31}ProfDev_j + u_{1j},$$

$$\text{where } \epsilon_{jt} \sim N(0, \sigma^2) \text{ and } var\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \begin{pmatrix} \tau_0^2 & \tau_{11} \\ \tau_{01} & \tau_1^2 \end{pmatrix},$$

$$\text{such that } u_{0j} \sim N(0, \tau_0^2) \text{ and } u_{1j} \sim N(0, \tau_1^2).$$

$CLASS_{jtc}$ is one of three domain-specific CLASS score outcomes (classroom organiza-

tion, emotional support, or instructional support) modeled as a function of the date of the video observation plus a set of teacher- and classroom-level covariates. The vector of covariates includes a dummy indicator for whether a teacher received a separate professional development treatment in phase one and another for whether they were newly added to the study in phase two. We also include dummy variable controls (fixed effects) for each city c to net out unique, time-invariant characteristics of the cities that might influence teacher-level average CLASS scores (e.g., geographic climate). We impute missing values of covariates with the mean for the analytic sample. We control for imputed covariates in all regression models.

Each variable in Equation 3.1 that corresponds to a season during the school year, $Fall_t, Winter_t$, and $Spring_t$, measures the submission date of a teaching video as the number of days between when the video was submitted from the start of that season up to its maximum length (divided by 30.25 to scale the variable to months). Values for $Fall_t$ range from 0.03 to 4.69 months; $Winter_t$ from 0 to 2.98 months, and $Spring_t$ from 0 to 3 months. We observe very few CLASS observations during the summer months when preschool programs tend to be closed. As a result, our $Spring_t$ measure includes a handful of summer observations (n=76 in June and n=2 in July out of 2,308 total video observations). Our results are not sensitive to their inclusion.

β_{0j} is the estimated intercept for the jth teacher (a random teacher intercept); γ_{00} is the average intercept for control teachers; and $\gamma_0 1$ is the increment to the intercept realized for treatment teachers. $u_0 j$ is a random intercept parameter that characterizes the remaining unexplained teacher-level variation around the overall treatment and control intercepts. β_{1j}, β_{2j}, and β_{3j} are teacher-specific CLASS score growth rates for the fall, winter, and spring, respectively. These vary as a function of professional development receipt, such that γ_{10}, γ_{20}

and γ_{30} represent average "business-as-usual" time trends in CLASS scores during the fall, winter, and spring, and the coefficients γ_{11}, γ_{21} and γ_{31} are the increments to these slopes for teachers receiving professional development.

We fit the model in Equation 3.1 by isolating the common random slope $u_1 j$ across seasons, indexing it with a separate time variable $Calendar_t$ that spans all calendar months and is centered around January 1, such that January 1 corresponds to $Calendar_t = 0$. This is equivalent to the presented model as $Calendar_t = Fall_t + Winter_t + Spring_t$, thus allowing us to factor out of the common random slope in the reduced form. Ultimately then, the model presented in Equation 3.1 fits a fixed seasonal curve (teachers' average rate of change in CLASS scores across fall, winter, and spring) adjusted and tilted by the random slope $u_1 j$ and random intercept $u_0 j$. We allow for the error variance structure of the random intercept $(u_0 j)$ and random slope $(u_1 j)$ parameters to be correlated.

Implications of CLASS Score Trends for Overall Teacher-Level Variability in Scores

Next, we explore how much of the overall variability in a teacher's CLASS scores, on average, can be explained by predictable time trends of growth and decline over the school year. To do this, we conduct a straightforward variance decomposition. First, we obtain a measure of the total within-teacher variability in CLASS scores, based on an unconditional multilevel model with no teacher-level covariates included. Second, we reestimate this same model but add estimates of average teacher time trends in CLASS scores and teacher-specific random effects around these trends. We obtain a revised estimate of the total within-teacher variability and, using this, assess the percent change from the initial estimate. The percent change reflects the proportion of the initial within-teacher variation explained by predictable growth and decline

in observed classroom quality over the year. The larger this proportion, the more meaningful CLASS score time trends are in explaining observed overall within-teacher variability in CLASS ratings. Appendix B provides further details on the estimation procedures behind this calculation.

Influence of CLASS Score Trends on Program-Level Accountability Outcomes

To further quantify the implications of documented within-year time trends in teacher CLASS scores, we use the Head Start Designation Renewal System as a case study. For most early education accountability systems, including the Head Start Designation Renewal System, quality evaluations involve measuring point-in-time CLASS performance of a sample of teachers. Then scores are aggregated to school- or agency-level averages.

We use our teacher-level findings on time trends to build a simulation model that predicts Head Start agency-level average CLASS scores at various time points during the school year. Details on the model are provided in Appendix C. Broadly speaking, for each week during the school year, we predict CLASS scores on the emotional support, classroom organization, and instructional support domains for eight random teachers in an "agency" using our time trend estimates. Our model also accounts for the likely correlation of teachers' CLASS scores within the same agency.[3] We run this prediction 1,000 times for each week during the school year and then compare the number of times (out of 1,000) that the simulated agency would pass its accountability review, based on Head Start accountability policy thresholds. We also account for different levels of underlying agency quality (e.g., above average, average,

[3]Ideally, instead of running a simulation, we would directly estimate growth and decline in Head Start agency-level average CLASS scores, given that accountability systems monitor agencies rather than individual teachers. However, our analytic sample does not include enough teachers within the same Head Start agency or preschool program to estimate trends at the agency/program level. We are unaware of any data source that contains multiple observations of teachers within a program assessed at multiple time points throughout the school year.

or below average) to pinpoint which types of agencies' accountability outcomes are most sensitive to within-year time trends in teaching quality.

We conduct simulations for two versions of the Head Start Designation Renewal System. Our main analysis relies on the absolute minimum CLASS thresholds currently in place. Any agency that scores below these thresholds is not eligible for automatic renewal of grant funding. Current thresholds are higher than those in place when the system was first implemented in 2012. The current thresholds require teachers' average scores be above 2.3 out of 7 on instructional support, 5 out of 7 on classroom organization, and 5 out of 7 on emotional support. Prior thresholds required average scores above 2 out of 7 on instructional support, 3 out of 7 on classroom organization, and 4 out of 7 on emotional support. We repeat the simulation using the prior CLASS thresholds to examine whether agencies' accountability outcomes across the school year look similar under the old policy.[4] We present those results in Appendix A.

3.7 Results

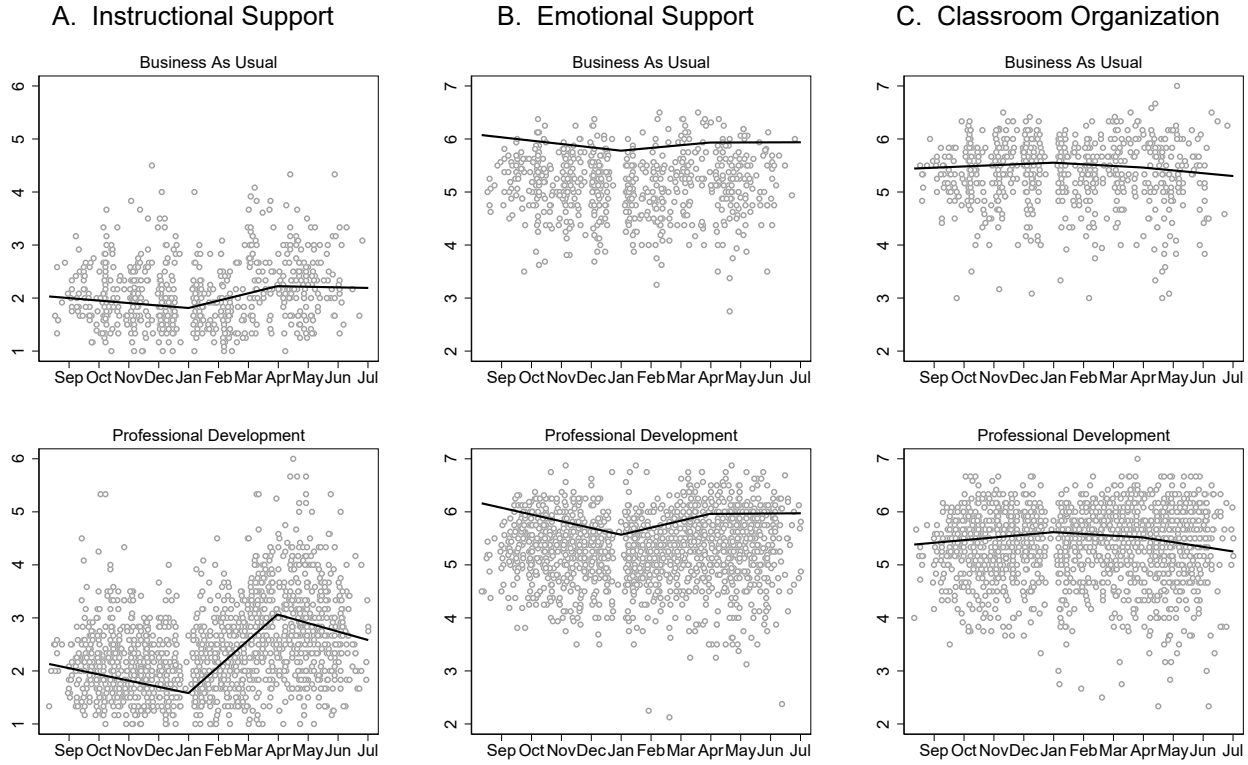
3.7.1 Trends in CLASS Scores over the Preschool Year

We first explored whether and how teachers' CLASS scores evolve during the school year across three measured domains of classroom quality. Table 3.1 presents the results from our multilevel linear spline models, and Figure 3.1 plots the results from each model. Models estimate the rate of growth in CLASS scores during the fall, winter, and spring months for

[4]Under the prior rules, it was possible for Head Start agencies that scored above the minimum CLASS thresholds to still fail their review, based on the CLASS. Any agency that scored in the bottom 10th percentile on any given CLASS domain in their review year would not receive renewed funding. Under the new policy, thresholds were increased and relative thresholds eliminated to ensure all agencies were held to the similar standards year to year.

our full sample of preschool teachers. We also estimate the increment to the growth rate for the subsample of teachers who received intensive professional development around classroom instruction under the larger intervention study. The models include random teacher intercepts and a single random slope parameter that characterizes teacher-specific variation around the fixed slope estimates.

Figure 3.1: Time Trends in CLASS Scores Over the Preschool Year, Multilevel Linear Piecewise Regression Results



Notes: Displays trends based on coefficients recovered from the multilevel piecewise linear regressions presented in Table 3.1, broken out by teachers receiving professional development and those teaching under business as usual. Raw data points shown as hollow grey dots.

Overall, we find a consistent pattern of growth and decline within the preschool year across two of the three domains of classroom quality. The quality of teachers' emotional and

Table 3.1: Results of Multilevel Linear Spline Models Estimating Trends in CLASS Scores

	(1) Emotional Support	(2) Classroom Organization	(3) Instructional Support
Fixed Slopes and Intercepts			
Fall months	-0.0631**	0.0237	-0.0468+
	(0.0222)	(0.0239)	(0.0267)
Winter months	0.0516+	-0.0306	0.1385***
	(0.0274)	(0.0296)	(0.0331)
Spring months	0.0017	-0.0529	-0.0111
	(0.0605)	(0.0653)	(0.0728)
Fall*Professional Development	-0.0009	0.0026	-0.0242
	(0.0271)	(0.0290)	(0.0326)
Winter*Professional Development	0.0289	0.0281	0.2211***
	(0.0324)	(0.0350)	(0.0392)
Spring*Professional Development	0.0005	0.0185	-0.1363+
	(0.0666)	(0.0719)	(0.0803)
Professional Development (0/1)	0.0898	-0.0599	0.1026
	(0.0930)	(0.0985)	(0.1080)
Constant	6.0759***	5.4420***	2.0322***
	(0.5368)	(0.5879)	(0.5293)
Random Slopes and Intercepts			
sd(months)	0.0385	0.0325	0.0421
	(0.0063)	(0.0085)	(0.0081)
sd(teacher)	0.3013	0.3334	0.2541
	(0.0193)	(0.0219)	(0.0197)
corr(months, teacher)	0.6599	0.9177	0.8328
	(0.1636)	(0.2551)	(0.1910)
Residual	0.4989	0.5431	0.6117
	(0.0083)	(0.0091)	(0.0100)
F-test: fall=0, winter=0, spring=0	0. 2376	0.0278*	0.0000***
F-test: fall*pd=0, winter*pd=0, spring*pd=0	0.6290	0.5040	0.0000***
Observations (n=CLASS ratings)	2,308	2,308	2,308

Notes: All models include covariates listed in Table 3.1 and control variables for teacher/program city, teacher receipt of professional development intervention under NCRECE Study phase one, and new teacher addition to the study in phase two. F-tests assess the null hypotheses that (1) fixed slopes for fall, winter, and spring for the full sample are jointly zero, and (2) the same coefficients for the subsample of teachers that received professional development are jointly zero. All models were estimated using restricted maximum likelihood estimation (REML). Standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$.

instructional support practices generally declines during the fall, from the start of the school year through the end of December. Then, for the middle portion of the school year, during the winter, observed teacher quality rises steadily. In the spring, beginning in April, teaching quality in emotional and instructional support plateaus, even declining slightly for those teachers who received professional development and who realized the largest improvements in the months prior. This pattern of findings suggests classroom quality in at least two of three measured CLASS domains follows a consistent curvilinear pattern of change within the school year in our sample. Observed teacher quality in classroom organization is largely flat across the school year.

This pattern of decline, growth, and plateau is most pronounced for observed instructional support teaching quality. Column 1 of Table 3.1 displays estimated instructional support growth rates within and across seasons for the fall, winter, and spring. Results suggest all teachers in the sample decline in instructional support quality during the fall at a rate of one-twentieth of a point on the CLASS (0.047 points) per month, though this finding is only marginally significant at conventional levels. Patterns in the fall do not differ, based on teachers' professional development receipt. Starting in the winter months, beginning January 1, teachers display increasingly better quality instructional support practices in the classroom, improving at a rate of one-seventh of a point (0.139 points) per month in their instructional support CLASS scores if not receiving professional development. Growth in instructional support plateaus in the spring, starting April 1, for these teachers not receiving professional development.

We find patterns of growth and decline are even larger in magnitude during the winter and spring months for teachers participating in professional development. During the winter, teachers participating in professional development improve at an *additional* rate of one-fifth

of a point (0.221 points) per month above and beyond teachers not participating. Total growth then for teachers engaged in professional development is roughly one-third of a point (0.360 points) per month in instructional support scores. During the spring, the gains that teachers undergoing professional development realized during the winter months in instructional support decline at a marginally significant rate of 0.136 points per month above and beyond overall trends in CLASS scores during the spring.

In terms of emotional support, we find similar trajectories of change in ratings within the preschool year, but these are smaller in magnitude and less precisely estimated. On average, teachers display modest declines in emotional support scores during the fall months. From the start of September through the end of December, emotional support ratings decline on average by a total of 0.25 points on the 7-point CLASS scale (calculated based on our estimate of a 0.063 point reduction in rated quality per month during the fall). Teachers' emotional support scores improve subsequently in the winter months, with a marginally significant increase of 0.052 points per month from January through March. These patterns hold across all teachers. We do not find evidence of any additional growth or decline in emotional support teaching quality among teachers receiving professional development.

In terms of classroom organization quality, we find no discernable time trends for the overall sample of teachers nor for the subset of teachers engaged in professional development training. The quality of classroom organization instructional practices appears largely consistent across the school year.

3.7.2 Implications of CLASS Score Trends for Overall Teacher-Level Variability in Scores

We then tested the extent to which within-year trends in teaching quality drive estimates of the total within-teacher variability in CLASS scores during the school year. Table 3.2 presents estimates of the within- and between-teacher variation in CLASS scores. Panel A shows results from unconditional models, including Intraclass Correlation Coefficient (ICC) estimates, which measure the proportion of the variation in CLASS scores due to differences *between* teachers. We find low ICC scores across each of the three measured CLASS domains. Instructional support scores show the lowest within-teacher consistency, with an unconditional ICC of 0.17. Emotional support and classroom organization unconditional ICCs are 0.32 and 0.31, respectively. These results indicate that most of the variability in CLASS scores comes from within-teacher variability or measurement error, not from differences between teachers' scores.

Next, we estimate a model that accounts for average and teacher-specific trends in CLASS scores to determine whether these variables help explain any of the total within-teacher variation in scores. We then calculate the percent change in the estimates of within-teacher variability across the two models. Results in Table 3.2, Panel B show that when we account for fixed- and teacher-specific time trends in instructional support ratings, we can explain 24% of the initially documented within-teacher variability in this domain. Put differently, nearly one-quarter of the variation in a teacher's instructional support ratings over the year is due to change over time in teaching quality, not idiosyncratic fluctuation or measurement error. In contrast, we find time trends explain little of the variability in teachers' emotional support and classroom organization ratings (6% and 2% of initial within-teacher variability, respectively) across the preschool year.

Table 3.2: Share of Within-Teacher Variation Explained by Average and Teacher-Specific Time Trends

	Emotional Support	Classroom Organization	Instructional Support
A. Unconditional model			
Within-teacher variation:	0.26	0.30	0.50
Between-teacher variation:	0.12	0.13	0.10
ICC:	0.32	0.31	0.17
B. Model with average + teacher-specific time trends			
Within-teacher variation:	0.25	0.30	0.38
Between-teacher random-intercept variation:	0.11	0.12	0.09
Between-teacher random-slope variation:	0.00	0.00	0.00
% change within-teacher variation (A to B)	*6%*	*2%*	*24%*

Notes: Displays estimates of the variability in CLASS scores across three separate specifications. Panel A, the unconditional model, shown as Equation 3.2, quantifies the degree of variation in a single teacher's CLASS scores and remaining variation between teachers without accounting for any teacher-level factors. ICC estimates measure the proportion of the variation in CLASS scores due to differences between teachers. Panel B, the model with average and teacher-specific time trends, shown as Equation 3.3, adds a teacher-specific random effect to the common fall, winter, and spring time trends. All specifications include fixed effects for each teacher's city location.

3.7.3 Influence of CLASS Score Trends on Program-Level Accountability Outcomes

Last, we explored how documented CLASS time trends might shape ECE programs' accountability outcomes, using the Head Start Designation Renewal System policy as a test case. Recall that under the Head Start Designation Renewal System, an agency is precluded from automatic funding renewal if average teacher CLASS scores fall below minimum thresholds on any one of the three CLASS domains. Recall, too, that programs are assessed at a single time point during the year and that these time points vary across programs.

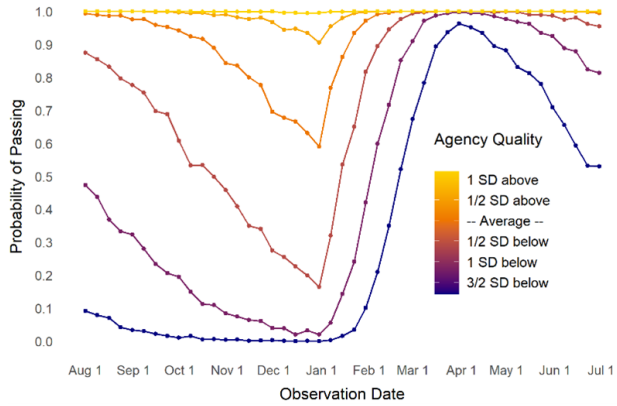
Broadly speaking, our simulations show that time trends in emotional support and in-

structional support give rise to different predicted accountability outcomes depending on when a Head Start agency is rated. Agencies with the greatest variation in predicted outcomes over the school year tend to be of average or below-average quality. These agencies are on the margin of the emotional or instructional support CLASS thresholds and thus susceptible to falling above or below the thresholds–passing or not–based on seasonal growth or decline in teacher ratings. Figure 3.2 shows results from all simulation analyses. Outcomes plotted in lighter colors (toward yellow) are among increasingly above-average Head Start agencies on the given domain. Outcomes plotted in darker colors (toward blue) are among increasingly below-average Head Start agencies on the given domain.[5]

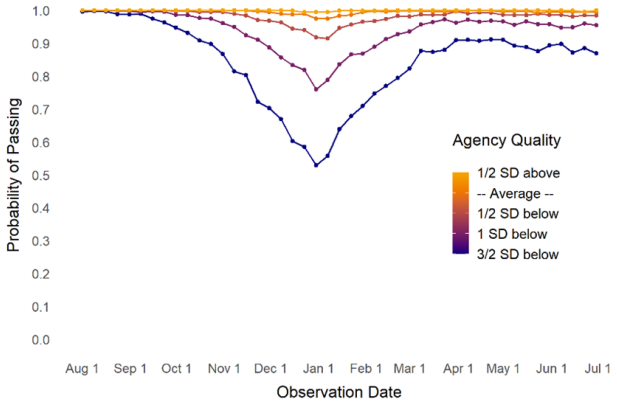
[5]We measure average quality for each CLASS domain using overall NCRECE sample averages; results are qualitatively similar if we rely on national Head Start average CLASS scores.

Figure 3.2: Simulated Head Start Agency Outcomes Over the Preschool Year Under Designation Renewal System CLASS Thresholds

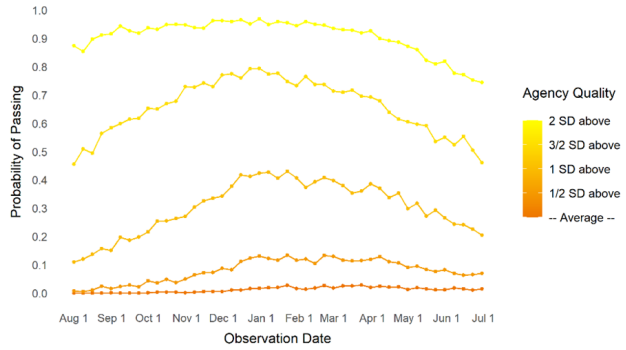
(a) Instructional Support, threshold=2.3



(b) Emotional Support, threshold=5



(c) Classroom Organization, threshold=5



Notes: Estimates are based on 1,000 simulations that each predict eight teachers' CLASS scores over a four-day "site-visit" period, run across 52 weeks of the year, based on the model shown in Equation 3.4. The Y-axis displays the proportion of a Head Start agency's scores above the Designation Renewal System threshold out of 1,000 simulations.

Head Start agencies' predicted accountability outcomes across observation dates vary most under minimum instructional support score requirements, shown in Figure 3.2a. Our simulation model predicts a Head Start agency 0.5 standard deviations below average in terms of baseline instructional support quality will pass its review (under at least this CLASS domain) 80% of the time if observed from September 1–4, but only 17% of the time if observed from January 1–4. The same agency has a nearly 100% chance of scoring above the instructional support threshold–thereby preventing their need to reapply for grant funding–if teachers are observed starting March 15–19 or later. In general, under the minimum threshold for instructional support scores, predicted accountability outcomes vary more widely the lower the agency falls in the instructional support quality distribution. Results imply that predictable growth and decline in instructional support ratings can put a range of Head Start agencies–especially those of average and below-average quality–at greater risk of failing their reviews, simply based on the time during the school year when they are evaluated.

Figures 3.2b and 3.2c show seasonal variation in predicted agency-level accountability outcomes under emotional support and classroom organization minimum score cutoffs. In terms of observed emotional support teaching quality, Head Start agencies 0.5–1.5 standard deviations below average face widely different likelihoods of failing their review, based on their observation date during the year. Our predictive model suggests these agencies are more likely to fall below the emotional support minimum score threshold and lose automatic renewed funding if evaluated during the winter months when teacher emotional support scores tend to decline to their lowest but not if evaluated in the fall or spring. Time trends in classroom organization do not substantially implicate Head Start agency accountability outcomes. We observe some variation in the likelihood of failing the review among above-average programs (see Figure 3.2c), but it is small in magnitude. Recall, we did not find

meaningful within-year growth or decline in teacher quality on this domain.

We also rerun our simulations using CLASS thresholds previously in place from 2012 to 2020 during the introduction of the Head Start Designation Renewal System, as opposed to the current CLASS thresholds (which constitute a revision since initial implementation). The previous thresholds were lower on all domains, meaning the bar for teacher quality was lower overall. We show simulation results based on these prior thresholds from initial policy implementation in Figure 3.A.2.

We find seasonal variation in CLASS scores matters much less for shaping accountability outcomes across the school year under the old thresholds. Fewer agencies' baseline levels of quality fall near the thresholds when they are set lower. Put differently, most Head Start agencies can score well above the thresholds. Therefore, growth or decline in teacher quality during the school year does not shift agency averages above or below the cut points to the same degree. These findings highlight that when policymakers raise minimum score thresholds, seasonal variation in teacher quality matters more. Agencies that previously outscored minimum required thresholds become closer–at baseline–to the revised higher cutoffs, which amplifies the importance of a slightly higher or lower score in any given month in the school year.

3.8 Discussion

The current study explores time trends in preschool teacher quality during the school year and resulting implications for quality measurement and ECE accountability systems. Using a flexible, multilevel modeling approach, we find small but meaningful patterns of growth and decline in instructional quality. Teachers' scores on two of three CLASS domains decline during the fall, improve during the middle of the school year, and plateau in the spring. These

seasonal patterns are most pronounced in teachers' instructional support ratings, that is, the extent to which instructional practices promote children's higher-order thinking. We find nearly one quarter (24%) of the overall variability or "noise" in teacher ratings of instructional support comes from predictable growth and decline in teaching quality, underscoring the importance of time trends in quality measurement. Last, we run simulation analyses to predict school-level accountability outcomes at various evaluation dates across the school year. We find sizeable variation in the probability that an agency passes or fails their accountability review under the Head Start Designation Renewal System based on the time of year teachers are assessed. In the fall and winter, lower quality agencies are more likely to score below CLASS thresholds. In the spring, almost all agencies have a high likelihood of passing.

Results from this study are consistent with prior research that generally finds teacher CLASS ratings improve over the course of the school year (cite). Where our results vary from other studies is in terms of the exact shape of the trajectory of change in teaching quality. For example, Cash and Pianta (2014) show instructional support quality peaks mid-year, while Meyer et al. (2011) find that instructional support ratings peak at the end of the school year. In contrast, we find scores—particularly instructional support scores—tend to peak in the spring and then level off for the rest of the school year. Notably, we are limited in our ability to shed light on the mechanisms behind observed growth and decline in CLASS ratings. Variation in these mechanisms across study contexts may underlie variation in the pattern of findings across studies.

Nonetheless, our time trend findings are in line with theory that suggests teachers focus more on classroom management initially during the school year—perhaps at the expense of emotional and instructional support practices—before turning to invest other domains of in-

structional quality Bohn et al. (2004) and Curby et al. (2011). Our results are also consistent with the idea that teachers have the most room for growth in terms of their instructional support scores. Teachers' instructional support ratings are consistently lower than their performance on other CLASS domains (Hamre et al., 2008, Office of Head Start, 2016, Perlman et al., 2016). This may motivate teachers to focus on performance improvement in instructional support practices more so than in other areas of instructional quality. Last, prior research shows especially large fluctuations across teaching occasions in instructional support CLASS scores relative to other class domains (Curby et al., 2011, Malmberg et al., 2010, Praetorius et al., 2014). Pronounced time trends in instructional support quality in our study align with this finding.

The evidence we uncover regarding time trends in CLASS ratings also raises questions about the optimal timing of accountability evaluations. If policymakers seek to identify the lowest quality programs, then evaluations in the fall and winter would most reliably identify these programs, based on our study's findings. In contrast, evaluations during the spring months when all programs perform at their best may prove more challenging for identifying poor performing programs. Even very low-quality programs were likely to pass their review in our simulation model if teachers were observed and rated in April, for example. On the other hand, if accountability systems seek to differentiate program performance across the entire quality distribution, including among higher quality programs, then our results suggest that January and February would be the best months to conduct CLASS ratings. During these months, teachers' rated performance is lowest, and thus evaluations more effectively distinguish even average quality programs from above-average programs.

Beyond raising questions about the optimal timing of accountability reviews, study findings indicate that current approaches to measuring ECE quality in high- and low-stakes

policy contexts may misclassify programs. We show ECE programs with similar underlying levels of quality that are observed and rated at different times during the school year face different probabilities of passing their reviews. ECE providers with quality levels close to set CLASS accountability thresholds are most susceptible to variation in accountability outcomes based on the timing of their assessment.

Recent changes to the Head Start Designation Renewal System may exacerbate this issue. In 2020, the Office of Head Start increased minimum required CLASS thresholds from 3 and 4 out of 7 for classroom organization and emotional support respectively to 5 out of 7 points for each domain. It raised the minimum threshold for instructional support scores from 2 to 2.3. (The instructional support threshold is scheduled to increase again to 2.5 in 2025.) With these changes, more Head Start agencies' average scores fall near the CLASS score cutoffs, making these agencies' likelihoods of passing or failing more susceptible to the influence of time trends in ratings. For example, in 2020, the average Head Start agency scored 2.94 on instructional support (Early Childhood Learning & Knowledge Center, 2020). Agencies below this average with quality levels close to the accountability threshold (2.3) may pass or fail simply due to the time of year teachers are rated.

The Head Start Designation Renewal System is one among many ECE accountability systems. Other ECE accountability systems, mainly state-level QRISs, tend to draw on multiple measures of ECE quality. For example, QRIS rating schemes may include average CLASS scores along with other quality metrics to jointly determine a program's quality rating or category. As a result, variation in the timing of CLASS evaluations may not have the same outsized effect on programs' accountability outcomes under QRISs as under the Designation Renewal System where CLASS scores alone can cause agencies to fail their review. In addition, QRISs impose fewer high-stakes consequences on poorly rated programs (though

poor QRIS ratings can drive programs out of business; Bassok et al., 2019). Therefore, variation in QRIS ratings due to variation in CLASS assessment timing may have fewer high-stakes consequences for programs.

Results from the current study should be interpreted with a few important caveats. First, our data come from a randomized control trial. Roughly half of the teachers in our sample were assigned to a professional development condition designed to improve the quality of their teaching practices. Among teachers receiving professional development, we find larger growth in instructional support CLASS ratings, though we find no differential growth or decline in other CLASS domains for teachers receiving professional development. As such, it is possible that our findings overstate the typical ECE teacher's growth in CLASS ratings during the year. However, we also note that many ECE teachers currently receive professional development to prepare for accountability reviews (Derrick-Mills et al., 2016).[6] Given this, our study sample (assessed in 2008 and 2009) may in fact align with more current ECE teacher experiences in the age of ECE accountability efforts.

Two additional limitations are worth noting. Study findings on time trends in CLASS ratings are estimated at the teacher- rather than school-level. Our simulation model aggregates up from teacher-level time trends to predict school-level accountability outcomes. However, an ideal analysis would rely on observations of the same teachers within the same schools at multiple points during the year to estimate growth and decline in school-level CLASS scores. To our knowledge, no such data exist to pursue this strategy, though future research efforts could focus on this approach. Additionally, our sample consists primarily of teachers working in publicly-funded preschools where most students in classrooms come from

[6]In a survey of 71 Head Start agencies, 100% of directors reported engaging staff in professional development on the CLASS or on child–teacher interactions more broadly in response to the 2011 implementation of the Head Start accountability policy (Derrick-Mills et al., 2016).

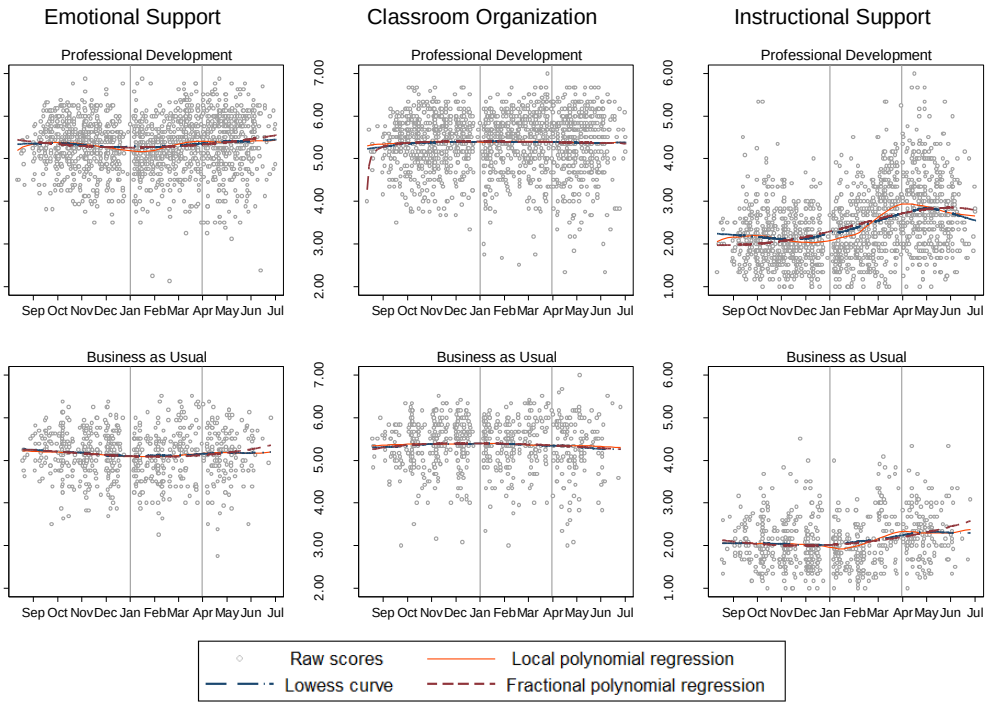
poor and near-poor households. To the extent that teacher quality evolves different in these settings, our findings may not generalize to preschools with different student composition or funding structures.

Despite study limitations, our findings on teacher-level time trends in CLASS ratings are relevant to research and policy efforts to measure ECE quality. First, our results suggest early education quality may not be static feature of classrooms. Second, our findings underscore the importance of measurement practices when it comes to monitoring ECE quality for accountability purposes. While current evaluation procedures for the Head Start Designation Renewal System and state-level QRISs call for in-person CLASS ratings, this requires evaluation teams spread out the timing of their visits to make them logistically feasible. Instead, accountability systems might supplement in-person ratings with video recordings of teacher instruction or opt exclusively for video observations. This shift would make it possible for all Head Start programs, for example, to be observed in the same week or month of the school year (albeit virtually).

Appendix

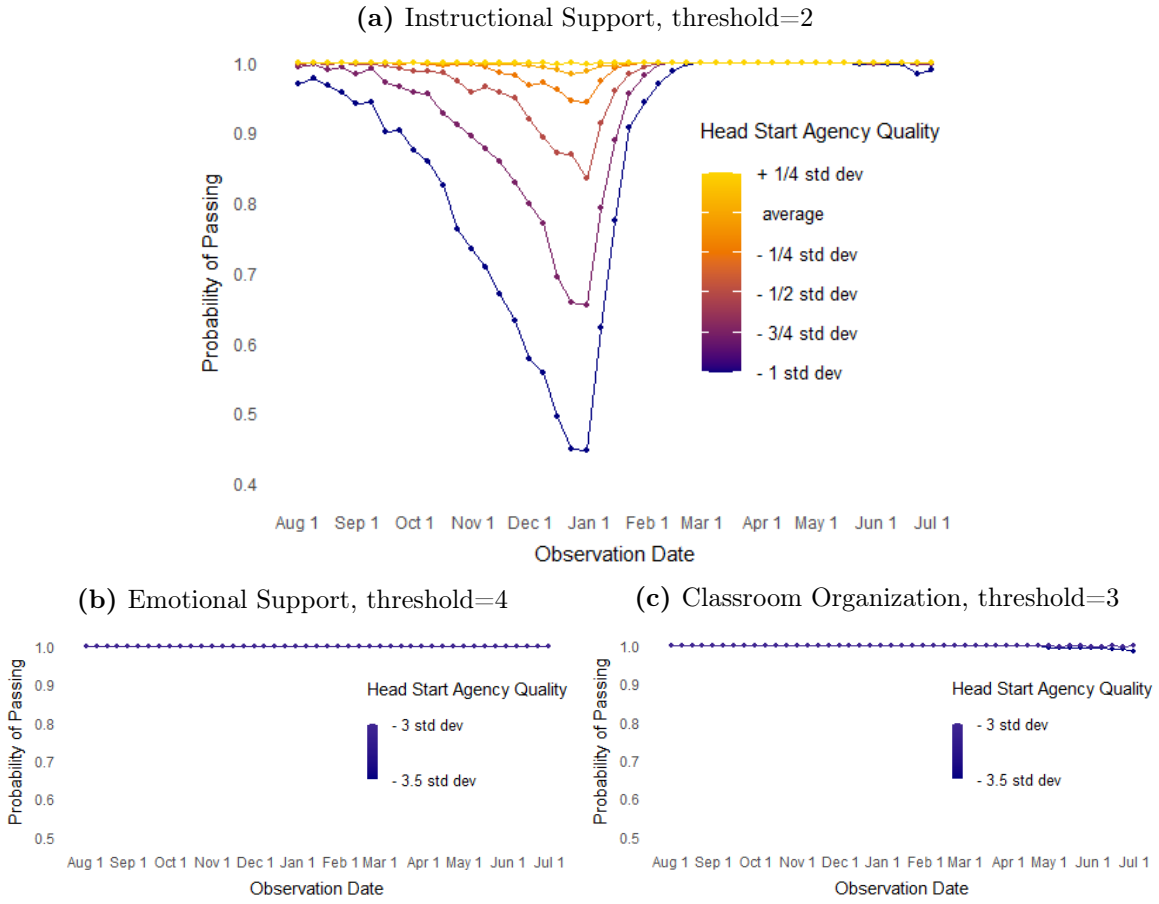
3.A Supplemental Figures and Tables

Figure 3.A.1: Flexible Models on CLASS Score Patterns Over the Preschool Year



Notes: Displays raw CLASS scores over the school year, overlaid with flexibly modeled local polynomial regression, locally weighted scatterplot smoothing (lowess) curve, and fractional polynomial regression approaches. Vertical lines are placed at Dec. 31 and Mar. 31, corresponding to the knot points used to construct linear spline models. The models shown characterize the shape of trajectories of change in CLASS scores over time but do not account for nesting of CLASS score observations within teachers.

Figure 3.A.2: Simulated Head Start Agency Outcomes Over the Preschool Year Under Prior Designation Renewal System CLASS Thresholds



Notes: Estimates are based on 1,000 simulations that each predict eight teachers' CLASS scores over a four-day "site-visit" period, run across 52 weeks of the year, based on the model shown in Equation 3.4. The Y-axis displays the proportion of a Head Start agency's scores above Designation Renewal System thresholds in place from 2012–2020 out of 1,000 simulations.

3.B Decomposing Total Within-teacher Variability to Identify the Share Explained by Time Trends

We estimate a series of models to assess the role time trends play in explaining total within-teacher variability in CLASS scores. To begin, we estimate the following unconditional multilevel model:

Level 1: Time (t)

$$CLASS_{jtc} = \beta_{0j} + \alpha_c + \epsilon_{jtc} \tag{3.2}$$

Level 2: Teacher (j)

$$\beta_{0j} = \gamma_{00} + u_{0j}, \text{where } \epsilon_{jt} \sim N(0, \sigma^2) \text{and } u_{0j} \sim N(0, \tau_0^2)$$

Equation 3.2 does not include teacher-level covariates because we are interested in estimating the *total* variation that exists within vs. between teachers in our sample. We include time-invariant city fixed effects α_c to parse out variation in CLASS scores between cities from variation attributable to teachers. Equation 3.2 produces an estimate of $\sigma^2 \stackrel{def}{=} var(\epsilon_{jtc})$, equivalent to the total unexplained within-teacher variation in CLASS scores. We also obtain $\tau^2 \stackrel{def}{=} var(u_{0j})$ $\tau_0^2 \stackrel{def}{=} var(u_{0j})$, the total variation between teachers in CLASS scores.

Next, we obtain a revised estimate of within-teacher variability, derived from the model below that accounts for average teacher time trends in CLASS scores within the school year and teacher-specific random effects around these trends:

Level 1: Time (t)

$$CLASS_{jtc} = \beta_{0j} + \beta_{1j} Fall_t + \beta_{2j} Winter_t + \beta_{3j} Spring_t + \epsilon_{jtc} \tag{3.3}$$

Level 2: Teacher (j)

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{1j}$$

$$\beta_{3j} = \gamma_{30} + u_{1j},$$

$$\text{where } \epsilon_{jt} \sim N(0, \sigma^2) \text{ and } var \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \begin{pmatrix} \tau_0^2 & \tau_{11} \\ \tau_{01} & \tau_1^2 \end{pmatrix},$$

$$\text{such that } u_{0j} \sim N(0, \tau_0^2) \text{ and } u_{1j} \sim N(0, \tau_1^2).$$

We calculate the difference between the unexplained within-teacher variation in CLASS scores σ^2 obtained from Equation 3.3, which accounts for average and teacher-specific trends in quality over time and the total unexplained within-teacher variation σ^2 obtained from the unconditional model, Equation 3.2. We divide this difference by the initial amount of unexplained within-teacher variation σ^2 from Equation 3.2. The resulting number estimates the proportion of initial within-teacher variation explained by predictable growth and decline in observed classroom quality over the year. The calculation is shown below:

$$\frac{\sigma^2_{Eq.3.2} - \sigma^2_{Eq.3.3}}{\sigma^2_{Eq.3.2}}$$

3.C Simulation Methods to Predict Head Start Agency Accountability Outcomes as a Function of Assessment Date

We begin our simulation approach by predicting a single teacher's CLASS domain score on a given date during the school year within a given Head Start agency. To do this, we use the following multilevel model:

Level 1: Time (t)

$$CLASS_{tij} = \pi_{0ij} + \pi_{1jk}Fall_t + \pi_{2jk}Winter_t + \pi_{3jk}Spring_t + \epsilon_{tjk} \tag{3.4}$$

Level 2: Teacher (j)

$$\pi_{0ij} = \beta_{00} + u_{0jk}$$

$$\pi_{1ij} = \beta_{10} + u_{1jk}$$

$$\pi_{2ij} = \beta_{20} + u_{1jk}$$

$$\pi_{3ij} = \beta_{30} + u_{1jk}$$

Level 3: Head Start Agency (k)

$$\beta_{00k} = \gamma_{000} + r_{00k},$$

$$\text{where } \epsilon_{tjk} \sim N(0, \sigma_y^2)\ u_{0j},\ \text{and } u_{1jk} \sim N(0, \sigma^2),$$

$$\text{such that } var\begin{pmatrix} u_{0jk} \\ u_{1jk} \end{pmatrix} = \begin{pmatrix} \tau_0^2 & \tau_{11} \\ \tau_{01} & \tau_1^2 \end{pmatrix}, \text{and } r_{00k} \sim N(0, \sigma^2_{agency}).$$

Random slope $(\pi_{1ij}, \pi_{2ij}, \pi_{3ij})$ and fixed slope estimates $(\beta_{10}, \beta_{20}, \beta_{30})$ for trends in CLASS

scores come from our NCRECE sample. To obtain these, we run a multilevel model similar to Equation 3.1. For simplicity, we include only one slope per time period ($\beta_{10}, \beta_{20}, \beta_{30}$), pooled across all teachers in the NCRECE sample, regardless of professional development receipt. The simulation model decomposes the random intercept for CLASS scores (π_{0ij}) into a teacher-specific component (u_{0ij}), drawn from our NCRECE sample, and a Head Start agency-specific component (β_{00j}).

The Head Start agency specific component of the model (β_{00j}) is made up of fixed agency quality (γ_{000}) and an agency-specific random effect (r_{00j}). We approximate average fixed agency quality (γ_{000}) as the average score from the 2,308 CLASS observations in our analytic sample.[7] We approximate agency-specific random effects (r_{00j}) using the assumed distribution, $r_{00j} \sim N(0, \sigma^2_{agency}$, where the measure of σ^2_{agency} comes from the 2009 FACES data set. The 2009 FACES data correspond most closely in time to our 2008-2009 NCRECE data on teachers that we use to predict time trends.

Our simulation approach follows the Head Start Designation Renewal System evaluation process as closely as possible. For example, the Head Start Designation Renewal System requires that external reviewers assess a minimum of eight randomly selected teachers within a Head Start agency (Office of Head Start, 2013)[8] As a result, we use the model above to simulate eight teachers' CLASS scores as an approximation of a single Head Start agency. Next, we define a site visit as a four-day period, based on Head Start reporting that the typical on-site visit lasts four to five days (Office of Head Start, 2013). For each of the eight

[7]We verify this approach is reasonable by comparing the NCRECE sample averages to the average Head Start agency in the nationally representative 2009 FACES Study. Averages for emotional support are 5.27 (NCRECE) and 5.31 (FACES). Averages for classroom organization are 5.38 (NCRECE) and 4.67 (FACES). Averages for instructional support are 2.33 (NCRECE) and 2.27 (FACES).

[8]The Office of Head Start uses a probability-driven algorithm to randomly select a sample of teachers from the Head Start agency for CLASS ratings and calibrates the number of teachers chosen to the size of the agency. No fewer than eight teachers can be rated (Office of Head Start, 2013).

teacher CLASS score predictions per agency, we randomly select one of the four days of the site-visit period to calculate the predicted CLASS score rating. We run this predictive model 1,000 times for each 4-day site visit. Each repetition draws teacher-specific and total unexplained variation at random from their estimated distributions and therefore builds in noise inherent in teachers' CLASS scores on a single site-visit date. For each of the 1,000 simulations, we then average the eight teachers' predicted CLASS scores. We use these averages to calculate how many times (out of 1,000) the simulated agency would pass its Designation Renewal System accountability review under the given CLASS domain. We repeat this prediction process for site-visit dates during every week of the preschool year. We also run the simulation for agencies at different assumed points on the quality distribution $N(0, \sigma^2_{agency})$. Finally, we plot the likelihood that a given agency of a given level of fixed quality (e.g., above average, average, and below average) passes its accountability review as a function of time over the school year.

Chapter 4

Targeted Public Pre-K and the Broader Child Care Landscape in Illinois

4.1 Introduction

Researchers and practitioners define early care and education (ECE) arrangements as settings where young children, ages 0 to 5 (not yet in Kindergarten), receive nonparental care. Reliable access to ECE supports both parent and child wellbeing. For parents, affordable child care is key to engaging in out-of-home work and maintaining economic self-sufficiency. For children, comprehensive and enriching ECE programs can improve kindergarten readiness, future educational trajectories, later-life health, and adult employment outcomes (Dynarski et al., 2013, Heckman, 2006, Ludwig and Miller, 2007, Thompson, 2018).

Recognizing the potential benefits of improved access to ECE, federal, state, and local governments have increasingly invested in public pre-Kindergarten (pre-K), designed to offer free early education services to preschool-aged children. Between 2011 and 2016, the federal government provided a total of $1.75 billion in grants to states through the Race to the Top—

Early Learning Challenge and Preschool Development Grants to support the implementation and expansion of public preschool (Congressional Research Service, 2016). More recently, under the American Families Plan, President Biden proposed a federal-level pre-K program for all children starting at age 3. Cities and states have increasingly spearheaded their own public pre-K initiatives as well. Currently, 62 public pre-K programs operate across 44 states (Friedman-Krauss et al., 2021).

Broadly speaking, policymakers who support public pre-K expansion do so with the goal of increasing the availability of pre-K services. Indeed, new funds may draw new providers into the market, thereby expanding the total amount of ECE services available to families. At the same time, however, public pre-K funds may go to already established ECE providers. If funds go to established providers, it is unclear whether these providers will expand their services in ways that improve ECE access. For example, new public pre-K funding could crowd out prior funding, such that providers serve the same number of children as before while being paid for by different funding streams.[1] In addition, receipt of public pre-K funding could cause providers to change the way they provide services and to whom, with positive or negative consequences for overall ECE availability. Finally, among preexisting providers, public pre-K funding could improve or reduce the odds that they stay in business. In this case, increased or decreased child care closures would shape overall "access" to care in terms of the number of child care providers available to families.

One reason public pre-K initiatives may be especially likely to fund already established providers is that—unlike new entrants to the market—established providers can leverage existing infrastructure, lowering their barriers to entry into the ECE market. For example,

[1]Outside of public pre-K funding, families may pay for care privately (i.e., pay tuition) or rely on other public sources of funding. Other sources of publicly funded care include (1) child care subsidies, offered by the state to defray the cost of child care for working parents with low incomes, and (2) Head Start, paid for by the federal government to provide early care and education services to poor and near-poor families.

existing ECE providers already meet licensing and regulation requirements (or have filed as exempt); have teachers, staff, personnel policies and recruitment strategies in place; and have facilities or experience locating buildings and service delivery spaces. In turn, these providers may be especially likely to host new public pre-K programs and may constitute a share of the policy response to public pre-K expansion.

Despite the importance of provider-level responses for the goals of public pre-K initiatives, little research exists on the first-order impacts of public pre-K expansion in the U.S. Limited ECE provider-level data makes it difficult to track outcomes at the ECE provider level. Moreover, public pre-K funding tends to be difficult to link with data on child care providers because public pre-K and child care records are generally housed in separate state agencies (education vs. human services). As a result, existing studies explore not how individual providers respond to state pre-K funding but rather how the aggregate stock of ECE services changes at the state (Bassok et al., 2014, Cascio and Schanzenbach, 2013, Fitzpatrick, 2010), county (Bassok et al., 2016), and neighborhood (Brown, 2019) levels in response to public pre-K expansion. Existing research therefore captures two distinct effects: (1) any immediate effects on providers that directly receive funding, and (2) any downstream competitive effects of the introduction of public pre-K among nonfunded providers – effects that result from increased competition in local markets due to a new public care option.

In this study, I explore possible direct effects of public pre-K expansion on ECE providers that receive funds. I focus on the context of Illinois' Preschool for All initiative, introduced in 2007. The initiative increased funding for half-day pre-K services and awarded funds annually through a competitive grant process. I match 3 years of state records on providers that received public pre-K funding during the initial program rollout obtained from the Illinois State Board of Education to a separate longitudinal database on the universe of

child care providers in Illinois. Using the longitudinal data on child care providers, I then identify the portion of providers that both received public pre-K funds and were new to service delivery vs. those that received public pre-K funds and were already established. Among the already established providers, I explore changes in their propensity to remain open, the number of children they can serve, the types of funding sources they receive, and other features of service delivery (e.g., infant/toddler care and number of distinct sessions) in response to public pre-K funding receipt.

My research design compares provider-level changes before and after public pre-K funding to changes over time in observably similar providers—identified using a propensity score matching approach—that do not receive public pre-K funding and do not operate nearby providers with public pre-K funding[2]. This strategy relies on time trends among matched nonrecipients to approximate trends in outcomes that would have occurred among public pre-K recipients in the absence of funding receipt. To interpret these estimates as causal, I must assume that no other unobserved sources of variation determine both the likelihood of public pre-K funding receipt and child care provider outcomes. I explore differences between pre-K funded and matched nonfunded providers *before* public pre-K funding expands to assess whether this assumption might be reasonably met; I find suggestive evidence to this extent.

Three main findings emerge. First, the large majority of ECE providers that receive state Preschool for All funding existed prior to public pre-K expansion. Put differently, Illinois' pre-K expansion funds go to very few new providers. Second, I find that public pre-K funded providers are more likely to remain in business during the 4 years after they receive funding.[3]

[2]I restrict possible nonfunded comparison providers to those in census tracts without a new publicly funded pre-K option. In this way, I minimize the likelihood that comparison providers are affected by nearby providers with public pre-K funding.

[3]I focus on public pre-K funded providers that existed at least 2 years before a public pre-K award. This

These results suggest public pre-K funding may keep providers that would have otherwise shut down from doing so, thereby increasing “access" to ECE by virtue of keeping these providers’ doors open. The magnitude of this effect is sizeable. Preschool for All funded providers show a 15 percentage point increase in the likelihood that they remain open 4 years after funding receipt. Third, among a balanced sample of providers that accounts for selective attrition, I find a marked increase in the number of distinct child care sessions for preschoolers offered by public pre-K funded providers in response to receiving funds. This change is closely aligned to the nature of Preschool for All funding that covers half-day rather than full-day ECE services. This finding suggests that while providers’ overall capacity to care for children in any given session did not change, total enrollment (i.e., the number of different children who come through the door each day) might have increased due to more separate sessions. I find no evidence of changes in provider-level capacity to enroll children or other types of services offered.

Findings on the immediate impacts of public pre-K on directly funded providers begin to inform our understanding of how the ECE sector responds to state pre-K expansion. However, an additional component not considered in the current analysis is whether pre-K funding creates competitive effects within the local ECE market. Competitive effects among nonfunded providers are possible, especially given that public pre-K funding shifts providers’ longevity and the number of distinct child care sessions they provide, both of which could influence the types of services nonfunded providers in the local market offer and to whom. Further, public pre-K services are free, which inherently creates competition among providers that charge fees.

---

allows me to match to nonfunded providers during a prepolicy time point. Some evidence suggests providers report on public pre-K funding starting the year before funds go into effect. This drives my choice to rely on the time point 2 years before funding to conduct the matching.

Prior research on the aggregate impacts of public pre-K programs speaks to the possibility of competitive effects of public pre-K on the child care market at large. However, past studies capture direct and competitive effects together, measuring outcomes among funded and nonfunded providers in nearby geographical areas. Broadly, they find increased public pre-K funding crowds out preexisting private child care providers, limiting the net expansion of early education opportunities (Bassok et al., 2014, 2016). Evidence also suggests increased availability of public pre-K comes at the expense of child care slots for younger, ineligible children (Brown, 2019, Bassok et al., 2014). However, these studies do not differentiate between changes that directly occur as result of public pre-K among funded providers vs. those that come about due to any competitive effects on nonfunded providers in the wider market.

In contrast to prior literature, the current study provides evidence on direct effects of public pre-K expansion in Illinois, where funding was designed to offer part-day services to low-income and otherwise at-risk families. These two program features—eligibility requirements and length of the preschool day–are likely to shape service provision among public pre-K funded providers in unique ways. For example, I find increased access in terms of number of distinct child care sessions, which is tightly linked to the part-day nature of the funding, rather than broad-based overall expansions in capacity. Thus, results will generalize best to public pre-K programs with similar program parameters, which are common across the U.S. More than half of state-funded pre-K programs (35 of 62) have an income requirement for eligibility, and nearly the same number (32 of 62) offer only part-day care (Friedman-Krauss et al., 2021).

4.2 Policy Background

4.2.1 Implementation of Preschool for All in Illinois

In July 2006, Illinois Governor Rod Blagojevich signed the Preschool for All initiative into law. The new legislation called for a universal pre-K program intended to provide all 3- and 4-year-old children with access to early education and came with an additional $45 million in funding in its 1st year (Illinois State Board of Education, 2007). The initiative built upon an existing public pre-K program designed for at-risk children, called the Pre-Kindergarten At-Risk Program (established in 1986). The goal of the Preschool for All initiative, then, was to expand eligibility beyond at-risk populations and eventually serve *all* children whose parents wanted them enrolled.

The initiative put Illinois on the map as a leader in early education services. The first full year of Preschool for All implementation occurred during the 2007–2008 school year.[4] That year, Illinois ranked 1st in terms of participation of 3-year-olds in state-funded pre-K and 11th in participation of 4-year-olds across the U.S. (Lloyd and Joseph, 2014).

Universal pre-K coverage was not immediate, however. Initial grants were awarded "based according to both need and the ranking of applications using a competitive grant process," (Illinois State Board of Education, 2007). Programs outside of Chicago were required to enroll at least 51% of children experiencing socioeconomic risk factors, such as low family income, homelessness, foster care participation, home environments where English was not the primary spoken language, or parental disadvantage, including teen parenthood or low parental education. The scoring criteria for these applicants included an assessment of (1)

[4]Preschool for All was signed into law in July 2006 and some programs were in fact funded in the 2006–2007 academic year. Limited data exist on grantees during this time period. Full implementation is often considered to have occurred in the 2007–2008 academic year.

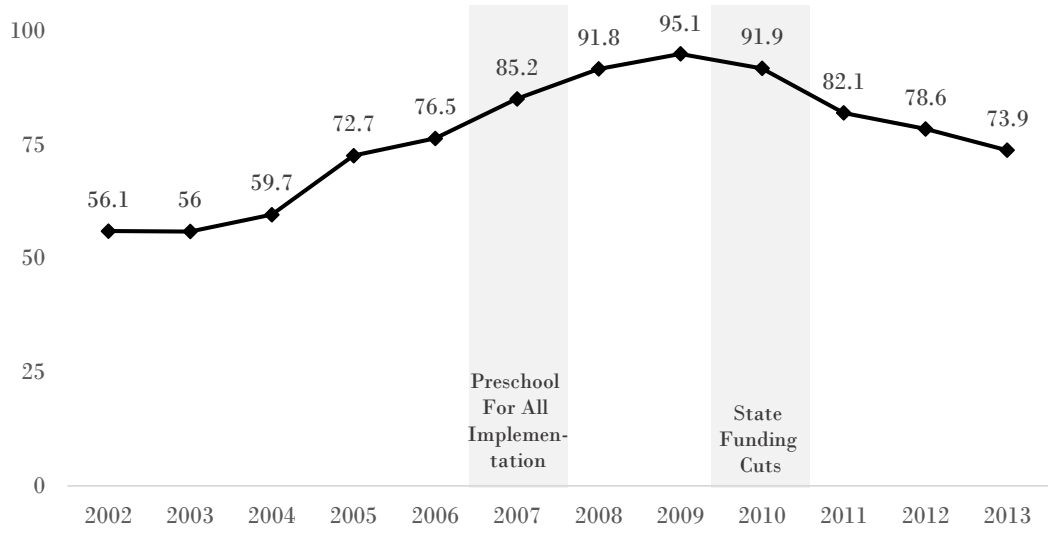
the need for preschool services based on current statistical, demographic, or descriptive information on the local community; (2) the quality of the proposed program, which is indicated by the adequacy of the screening process to identify children who are at-risk, the appropriateness of the proposed education program, the description of proposed family involvement services, and more; (3) the experience and qualifications of staff, based on the applicant's proposed plan for meeting staff education and professional development requirements; and (4) the accuracy, consistency, and cost-effectiveness of the proposed budget (Burgett et al., 2018).[5] For programs located in Chicago, the public school system was allocated a portion of Preschool for All funds and awarded funds based in part on the share of students receiving free and reduced-price lunch (Thomas et al., 2011).

Almost any public or private ECE provider could apply to offer Preschool for All. Eligible applicants included: public school districts, university laboratory schools, child care centers, regional offices of education, charter schools, community colleges, community organizations, private preschools, park districts, faith-based organizations, home-based child care networks, and other settings. If not part of a school district, applicants needed to provide evidence of their capacity to provide early childhood education services, including the agency's mission statement, organizational structure, and accreditation.

Illinois granted new Preschool for All awards in the first 3 years of the program, fiscal years (FY) 2007–2008, 2008–2009, and 2009–2010. Applicants were required to devote more than 50% of proposed slots to socioeconomically at-risk children. Planned expansion beyond this priority group met challenges when the Great Recession took effect. The State of Illinois cut funding for existing Preschool for All grantees by 10% in 2009–2010. In the next year

[5]I have gathered data on grant applicants from 2008 and 2009. Data include application scores and the name of the applicant. They do not include applicant address. Future analyses may leverage application data if it is possible to identify the addresses of each applicant to link to child care markets.

Figure 4.1: Number of Children Served in State-Funded Preschool in Illinois (in 1,000s)



Notes: Data come from the Illinois State Board of Education, presented in Lloyd and Joseph (2014). Enrollment reflects the number of children in *any* publicly funded pre-K program, including both Preschool for All and programs that pre-date the Preschool for All initiative. 2013 enrollment numbers are estimates.

(FY2011), no new awards were given and the State implemented a new requirement that at least 80% of slots in Preschool for All programs be reserved for at-risk children. With these changes and reduced funding, a number of existing Preschool for All grantees opted out of the program in the 2010-11 school year (Lloyd and Joseph, 2014). In the years that ensued, FY2012 to FY2017, no Preschool for All grants were given to new applicants. New awards began again in FY2018.

Figure 4.1 shows aggregate trends in enrollment in all state-funded pre-K programs in Illinois. Increases in public pre-K enrollment correspond with the period when Preschool for All awards began, starting in 2007. Enrollment began to taper in 2010 at the same time as statewide budget cuts went into effect.

4.2.2 Provision of Preschool for All Services

At the time of implementation, Preschool for All was rated as a relatively high-quality early education program, meeting 9 of 10 quality benchmarks set by the National Institute for Early Education Research (Barnett et al., 0007). From program inception through the present day, Preschool for All funding has covered half-day care, guaranteeing children a minimum of 2.5 hours of preschool services per day for 5 days per week during the academic school year. Providers can blend Preschool for All funding with other sources of public and private funding to offer more comprehensive services, which state-level guidance encourages. A primer for prospective providers released in 2007 suggests applicants might use collaborative models such as embedding Preschool for All programs in full-day, full-year child care programs or combining Preschool for All with Head Start funding to serve children eligible for Head Start who also meet Preschool for All socioeconomic risk criteria (Illinois State Board of Education, 2007). This same primer, however, stipulated that "Funds provided for Preschool for All must supplement, not supplant, funds received from other sources for the same purpose" (Illinois State Board of Education, 2007), underscoring the policy goal to expand preschool services without crowding out existing services that providers already offer.

Preschool for All grants go into effect starting July 1 or later and run through June 30. Subsequent funding is contingent on the availability of state funds and satisfactory progress in the preceding grant year. Grantee recipients can and often do distribute funds to smaller ECE sites that then offer Preschool for All services directly to families.

4.2.3 Child Care Landscape in Illinois

The child care landscape as a whole includes all settings where young children—ages 0 to 5 and not yet in Kindergarten—receive nonparental care. The market for this type of care consists of a diverse set of public and private options. Private ECE providers include informal child care arrangements with friends, relatives, or nannies, as well as more formal settings such as family child care homes and child care centers. While informal providers are an important source of child care, they do not generally constitute the "stock" of care available to all families; many of these arrangements occur through family and friend networks. Additionally, informal providers are not governed by state licensing requirements or regulations, limiting available data. As a result, in this paper, I focus on more formal care arrangements, group child care settings offered in both families' homes and child care centers.

Among home-based providers, Illinois' licensing requirements stipulate that if an individual cares for more than 3 children, including their own children under age 12, the home must be licensed as a family child care provider. Family child care providers can serve no more than 8 children under 12 years old, with a maximum of 2 children under 30 months (roughly 2.5 years old) and a maximum of 6 children under age 5. If an ECE program operates in a center-based facility, it too must be licensed through the state. Exceptions include child care centers operated as part of a public or private elementary school, college, or religious institution, which qualify as license-exempt. Because many Preschool for All sites are located in schools, many operate as license-exempt programs.

Other public ECE options in Illinois include the Head Start program, a federally funded early education program for families living in poverty; and child care subsidies, funded by the federal Child Care and Development Fund to cover some or all child care costs for low-income working parents. Head Start and its complement for younger children, Early Head

Start, provide full-day ECE services free of charge to families with children ages 0 to 5 years old. The child care subsidy program is not entirely free of charge. It requires parents pay a co-pay and covers full-time care for children ages 0 to 5 years old, as well as subsidized before-/after-school care for children up to 13 years old.

When implemented, public pre-K programs join the wider set of ECE options families have at their disposal for securing nonparental child care arrangements.

4.3 Related Literature

A small literature explores the introduction and expansion of publicly funded pre-K and its corresponding influence on the ECE sector. Broadly speaking, studies show universal pre-K programs increase overall preschool attendance, both public and private, (Cascio and Schanzenbach, 2013, Fitzpatrick, 2010) and grow the size of the ECE sector (Bassok et al., 2014, 2016). However, increases in the number of ECE options due to public pre-K are partially offset by reductions in the total number and capacity of child care providers, measured at the state (Bassok et al., 2014), county (Bassok et al., 2016), and neighborhood (Brown, 2019) levels. Limited net expansion suggests a majority of public pre-K slots—roughly 60% in Georgia and 75% in Florida—were offered in preexisting provider settings (Bassok et al., 2014, 2016).

Additional research explores spillover effects of universal public pre-K programs on the availability of care for younger, ineligible children. Bassok et al. (2016) find fewer 3-year-olds enrolled in any form of child care after the implementation of public pre-K in Florida, and Brown (2019) documents decreases in private center-based capacity for children ages 2 and younger in New York City in neighborhoods with more universal pre-K sites. Head Start programs appear to absorb some of this decline in private market ECE slots for children

too young to attend preschool, highlighting the potential for market-wide adjustments in response to public pre-K programs. Specifically, Bassok (2012) finds that Head Start programs serve a greater share of children 3 years old and under in response to nearby public pre-K expansion. Some evidence also suggests that states may face an insufficient number of ECE staff after public pre-K is implemented (Bassok et al., 2014), and the quality of programs serving younger children—as proxied by inspection violations—may decline (Brown, 2019).

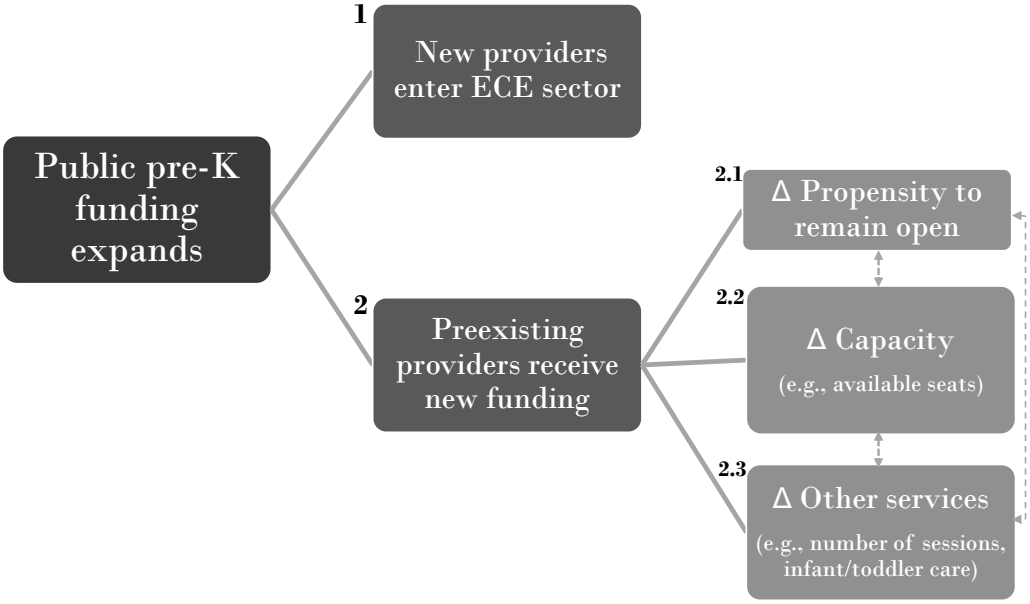
While current research begins to shed light on the impacts of universal pre-K on the broader ECE landscape, gaps remain. First, within this literature, it is unclear exactly which providers expand, contract, or go out of business and whether these responses come from direct receipt of funding or increased competition in the market at large. For example, reductions in the size of the total ECE sector could stem from preexisting providers repurposing themselves into public pre-K providers, or from nonfunded providers scaling back capacity to increase quality or going out of business in the face of increased competition. Additionally, the majority of studies focus on universal pre-K programs, which all families are eligible to attend regardless of income. These programs may attract different types of providers to apply for funds or enter the market, thereby influencing the nature of both direct effects on pre-K funded providers and competitive effects among nonfunded providers.

In contrast to prior research on the influence of public pre-K expansion on the total stock of ECE options available in the market, I focus on changes in provider-level ECE service provision among providers that directly receive public pre-K funding, separate from nonfunded providers that may face competitive pressure from direct recipients. In addition, I focus on how a half-day public preschool program that serves low-income or at-risk preschool-aged children interacts with the existing ECE landscape, generating evidence on a type of public pre-K program yet to be studied.

4.4 Conceptual Framework

In this section I outline potential consequences of expanded public pre-K in Illinois, focused on pathways through which increased funding may have direct effects on Preschool for All providers themselves. These are shown graphically in the conceptual model in Figure 4.1. Recall that in Illinois, Preschool for All funding enables providers to offer half-day care to children from predominantly low-income or otherwise at-risk families. Also recall that almost any child care or education provider in Illinois can offer Preschool for All funded services. I end the section by discussing possible competitive effects of public pre-K funding on ECE providers who do not receive the funding themselves but potentially compete with newly funded sites.

Figure 4.1: Conceptual Framework



As shown in boxes 1 and 2 of Figure 4.1, public pre-K expansion can either induce new educational or ECE facilities to open or provide new funding to preexisting providers that

then begin to offer public pre-K services. Barriers to opening a new child care center or home may limit the number of new providers drawn into the market in response to public pre-K funding. In Illinois, the process involves locating a building or service delivery space, making necessary renovations to meet licensing and regulatory requirements, obtaining a license, and planning the program (e.g., setting objectives, recruiting staff, writing personnel policies, establishing age ranges served, operating hours, etc.). As described in a primer for potential new center-based providers, it is a process that can be "overwhelming with all of the procedures," (Illinois Action for Children, 2021).

Beyond barriers to entry in the ECE market, features of the public pre-K program itself may shape whether new or preexisting providers opt to offer public pre-K. Important features may include which types of providers can apply to offer public pre-K (e.g., public schools, private centers, family child care homes, etc.), the length of the funded school day, the types of children eligible to attend, the per-child funding rate, and the administrative burden of the application. In Illinois' context, public pre-K expansion, designed to serve socioeconomically disadvantaged communities, may constrain potential new providers' profitability. Additionally, Preschool for All funds provide only half-day preschool services and public guidance encouraged collaboration with other existing providers (Illinois State Board of Education, 2007). As such, in Illinois, existing providers may have been best incentivized to offer Preschool for All.

Among preexisting ECE providers, a new source of revenue in the form of public pre-K funding could directly impact providers in a few key ways. First, as the box labeled 2.1 in Figure 4.1 shows, public pre-K funding could shift the likelihood that ECE providers remain in business. For example, the distribution of public funding could be unreliable, frequently delayed, or subject to budget cuts, possibly destabilizing providers' finances and leading

to closures. Similarly, it might crowd out other important sources of revenue that help provider stay afloat. Alternatively, public pre-K funding could provide a steady revenue stream relative to other funding sources, thus buoying provider finances and preventing closures. The direction of impacts of public pre-K funding on provider attrition may depend on context. Shortly after the start of Preschool for All expansion in Illinois, for instance, the Great Recession began. Therefore, relative to other providers that perhaps relied more on private-pay revenue, providers with Preschool for All funding may have remained more financially viable.

Second, as the box labeled 2.2 in Figure 4.1 reflects, preexisting providers that receive public pre-K funding may adjust the number of child care slots they offer in response. On one hand, with the new revenue source, providers will need to offer new public pre-K services. With increased funding, they may be able to hire new teachers and expand operations, adding onto preexisting ECE options offered. With time, public pre-K funding could also *crowd in* additional sources of revenue. Parents whose children are enrolled in Preschool for All might pay privately for extended hours of care to supplement half-day public pre-K services, or public pre-K dollars could pave the way for providers to participate in other types of publicly funded programs (e.g., Head Start or child care subsidies), possibly further expanding overall capacity. On the other hand, providers may face constraints, such as limited numbers of available classrooms or tight teacher labor markets. In this case, public pre-K dollars could crowd out other prior sources of revenue, causing no change in capacity.

Finally, new public pre-K funding could shift the types of services preexisting providers offer and to whom, as depicted in box 2.3 in Figure 4.1. To begin, the Preschool for All program explicitly funds only half-day public pre-K services. This funding could therefore incentivize providers to add more half-day sessions to their daily schedule, either by ex-

panding their hours of operation or converting some full-day services into part-day options. Additionally, providers might have new incentives to convert child care slots—previously set aside for infants and toddlers or school-aged children—into public pre-K slots. Notably, ECE providers generally do not profit off of infant/toddler child care services due to the high teacher-child ratios required. Instead, providers commonly use profits from 3- and 4-year-old care to subsidize the costs of providing infant/toddler care, planning to recoup costs as infants and toddlers grow older in the program (Brown, 2018; Stoney, 2015). This funding structure makes it potentially lucrative for preexisting to ECE providers to repurpose infant/toddler slots into public pre-K ones. At the same time, licensing regulations can make shifts in service provision across age groups difficult, potentially limiting this type of response to pre-K funding.

Beyond direct effects on funded providers (shown in Figure 4.1), public pre-K may impact nonfunded providers in the local market for ECE (not shown in Figure 4.1). nonfunded providers that compete with public pre-K programs may adjust prices, quality or types of services offered in response to public pre-K expansion. However, the scope for competitive effects likely depends on whether public pre-K expansion brings new providers into the ECE market or funds preexisting providers. If the latter, whether and how preexisting providers respond to new pre-K funding will then shape the nature of the market-wide competitive effects. In the current paper, I focus on the initial impact of public pre-K expansion in Illinois on funded providers themselves. Initial impacts, however, can help shed light on the potential for downstream competitive effects of public pre-K on the market at large.

4.5 Research Questions

To test the pathways laid out in the conceptual framework, I address the following research questions:

1. What proportion of public pre-K funded providers in Illinois are new to the ECE sector and what proportion already existed?

2. Does public pre-K funding receipt affect preexisting ECE providers' operations by

   (a) changing the likelihood that providers remain in business,

   (b) expanding providers' overall or pre-K-specific capacity to enrol children, or

   (c) altering other types of ECE services (i.e., number of sessions per day or provision of other types of publicly funded services) or to whom they offer services (i.e., availability of infant/toddler or school-age care)?

4.6 Data

4.6.1 Public Pre-K Data

I obtain data on Preschool for All expansion through a Freedom of Information Act request submitted to the Illinois State Board of Education (ISBE). ISBE records contain the grant number, grantee name, service-provision site name, service-provision site type (licensed or exempt), and addresses of all Preschool for All funded sites from FY2008 to FY2017. Importantly, this allows me to identify the ECE provider locations where children actually receive Preschool for All services, not just the address of the overall grant-recipient agency that received the funding award. While grantees can directly administer Preschool for All, they can

also provide funding to other ECE providers to offer public pre-K services. I am interested in the service-provision site level, which allows me to identify individual ECE providers.

Unfortunately, ISBE records do not include site-level information for FY2007 grants, those awarded during the 1st year of program implementation. ISBE did not maintain site-specific records in this 1st year of public pre-K expansion. Therefore, I do not include Preschool for All providers first funded in FY2007 in the analysis.[6] Additionally, ISBE records do not include Preschool for All providers funded in the City of Chicago, given that Chicago Public Schools oversee Chicago-based public pre-K services. Finally, budget cuts led the State of Illinois to freeze new Preschool for All grant funding between FY2011 and FY2017. [7] Therefore, I focus public pre-K expansion through new Preschool for All grants awarded between FY2008 and FY2010.

4.6.2 Child Care Provider Data

I link Preschool for All records to longitudinal data on all child care providers in Illinois. Data on child care providers come from the Illinois Network of Child Care Resource and Referral Agencies (INCCRRA). INCCRRA collects, maintains, and provides this data to state government officials, researchers, community planners, and human service organizations for the purposes of identifying and addressing the child care needs of children and their families in Illinois and does so under the contract and authority of the Illinois Department of Human Services.

[6] I can observe sites that were funded by a grant first awarded in FY2007 starting as of FY2008. However, I cannot be sure the FY2008 sites I observe are the same ones that were first funded in FY2007. Therefore, to be conservative in my analysis, I exclude data on any sites where the grant funding was first awarded in FY2007.

[7] In FY2012, the State did grant a limited number of Preschool for All awards to prior grant recipients that had lost or relinquished initial funding. I do not separately explore FY2012 reinstated grant funding. I focus instead on the 1st year a provider received Preschool for All funding regardless of whether they continued to maintain the funding.

Data from INCCRRA provide yearly, address-level records on the universe of licensed and license-exempt child care centers and licensed family child care homes in Illinois from FY2004 through FY2014. Data include characteristics of child care providers, including provider type, age ranges served, hours of operation, number of sessions, self-reported sources of public funding (Head Start, state pre-K, or whether they accept child care subsidies)[8], and quality ratings (when available).[9] Center-based providers report on their licensed capacity to serve various numbers of children across different age groups, while family child care homes report enrollment and vacancies by age group. The data also include prices charged by age group and by session type (e.g., full-day vs. overnight). All data are measured as of June 30 of the given year, which aligns with the end of the Illinois fiscal year as well as the end of annual Preschool for All grants.

4.6.3 Neighborhood Demographic Data

Last, I draw on data at the census-tract level to characterize neighborhoods where child care providers are located. Data come from (1) the 2000 Decennial Census, and (2) the American Community Survey (ACS) 2005–2009 5-year estimates. Data from the 2000 Decennial Census provide a snapshot of neighborhood characteristics before Preschool for All grants are awarded; it is the most recently available data at the tract-level collected before FY2007. However, the Decennial Census includes fewer detailed demographic indicators than the ACS. As such, I also draw on the ACS. However, the earliest 5-year estimates (2005–2009) include some data gathered after the implementation of Preschool for All in 2007.

[8]Provider reported funding is sometimes inaccurate or incomplete, rendering these data only partially reliable.

[9]Quality ratings come from Illinois' Quality Rating and Improvement System, which was first implemented in 2008. Providers are not required to participate in the rating system. All licensed providers are given the lowest rating and then can elect to demonstrate additional quality features and earn higher ratings. Ratings remain in place for 3 years before they are again reviewed and updated.

I use data on child care provider addresses to geocode the spatial location of all ECE providers (both public pre-K funded and nonfunded) and locate them within their corresponding census tract. Census tracts approximate the size of a neighborhood. They include roughly 4,000 residents on average and range in population from 1,200 to 8,000 residents. Census tracts are commonly used to define ECE markets (Jessen-Howard et al., 2018, Malik et al., 2018)—the area within which families look for care. Therefore, they provide a meaningful measure of neighborhoods where providers operate.

4.6.4 Matching Procedure

To identify Preschool for All providers within the INCCRRA data, I use data on addresses of Preschool for All providers from the ISBE data and match them to addresses of providers listed in the INCCRRA database. I rely on a fuzzy matching procedure that uses street address, city, zip code, county, and geocoded latitude/longitude coordinates to identify likely matches and hand-code any matches that fall below 100% odds. To increase the likelihood that I identify corresponding records, I allow Preschool for All providers to match to child care provider addresses recorded in *any* year of the longitudinal dataset. Then, I use the unique provider ID of the child care match to align Preschool for All records and child care records in the precise year of public pre-K funding receipt. For all unmatched Preschool for All providers, I identify their city location and hand-review the corresponding INCCRRA data on child care providers in that city to ensure all possible matches are detected.

Of 227 distinct Preschool for All funded providers from grants between FY2008-FY2010, I match 84% to corresponding entries in the INCCRRA database. INCCRRA gathers data on Preschool for All sites through regular data collection from regional Preschool for All

grantees and does not have a 100% response rate.[10] Therefore, the fact that I identify 84% rather than 100% of Preschool for All providers in the INCCRRA data is not unexpected. Of the unmatched Preschool for All sites, more than 80% are license-exempt and located in elementary schools based on documentation provided with the funding records. This corresponds to challenges INCCRRA reports facing when gathering information on license-exempt child care providers for whom reporting is not automatically required. (Note, public schools operate as license-exempt ECE providers.)

4.7 Empirical Strategy

I rely on Preschool for All sites matched to longitudinal child care provider records from INCCRRA to characterize whether public pre-K funded sites are in fact new to service delivery in the ECE sector. I define "new" providers as those that did not exist within the INCCRRA database until their 1st year of Preschool for All funding. I also capture "recently opened" providers as those operating only for 2 years or less before receiving public pre-K funding.[11]

Next, I focus on established providers that receive Preschool for All funding. I characterize differences between these providers and other nonfunded providers in the ECE sector, separately for center-based and family child care providers. I present information on self-reported sources of funding, types of service delivery (e.g., any infant/toddler or school-aged

[10] Regional Child Care Resource and Referral agencies throughout Illinois are tasked with calling local Preschool for All grantees to gather information on new providers. The regional Child Care Resource and Referral agencies report data back to INCCRRA, which then feeds back into the state-wide database. INCCRRA has confirmed that Preschool for All providers may be missing from the INCCRRA data, especially license-exempt providers for whom reporting is not automatically required.

[11] As discussed, there are a set of Preschool for All providers that do not ever match to a child care provider in the INCCRRA database. These may constitute "new" providers but may not be captured as such using the definition described here.

care), average total capacity and age-specific capacity (broken out by infants/toddlers vs. preschoolers), and the average number of sessions per day by age group. I then present baseline characteristics of neighborhoods with and without new public pre-K funded services under Preschool for All expansion.

Observable differences between funded and nonfunded providers highlight the challenge in identifying the impact of public pre-K funding on ECE providers. Put simply, providers do not receive Preschool for All funding at random. Factors that predict funding receipt may also drive changes in future outcomes for these providers. In the absence of random variation from an experimental or quasi-experimental framework, I rely on a propensity score matching technique to explore the impact of public pre-K funding on providers. My approach compares changes over time among ECE providers that receive Preschool for All funding to changes over time among observably similar providers that do not receive funding. I assume that the matched comparison group provides a valid estimate of the counterfactual outcomes for public pre-K recipients had they not received funding. For this assumption to hold, there must be no other *unobserved* factors that impact both the likelihood of public pre-K receipt and future outcomes. This is a strong assumption, given that I cannot observe and account for all factors that might lead a provider to receive Preschool for All funds. Therefore, I interpret my results as initial evidence on how public pre-K funding implicates preexisting ECE providers' operations.

I match funded and nonfunded (comparison) providers two years before Preschool for All funding begins for a given provider. My choice to match providers *2* years prior to Preschool for All receipt—rather than one year prior—stems from descriptive evidence that providers report state pre-K-related changes to ECE services as soon as they know of their award, 1 year prior to the actual start of funds (see Table 4.2). As a result, I restrict

my analysis to ECE providers open 2 years or more before receiving public pre-K funds. Because there is scope for competitive effects on nonfunded providers, I also limit possible comparison providers to those in census tracts where no other providers received Preschool for All funding.

I require an exact match on provider type (center vs. family child care). I also require that comparison providers match to Preschool for All providers in the exact calendar year that corresponds to 2 years before Preschool for All funding receipt. After this exact match, I estimate a provider's propensity to receive public pre-K funding using a host of provider- and neighborhood-based variables, presented in Table 4.A. As shown in Table 4.A, I create two matched samples. The first includes all providers, any Preschool for All provider matched to any comparison provider at $t = -2$ regardless of how long each remains in operation. The second restricts to a balanced sample of ECE providers, requiring both Preschool for All providers and matched comparisons to remain present in the data for the entire study window, 4 years before and 4 years after public pre-K receipt. This latter sample constitutes a smaller group of providers and requires matching on a different combination of variables to optimize the match.

I use nearest-neighbor matching with replacement to identify the 10 most observably similar comparison providers for each Preschool for All provider. I then run flexible event study regressions with fixed effects for each match group, consisting of one Preschool for All provider and 10 matched comparisons. My model takes the following form:

$$Y_{pgk} = \sum_{k \neq -2} \mathbb{1}(t = t_i^* + k)\beta_k + \phi_k + \alpha_g + \varepsilon_{igt} \tag{4.1}$$

where β_k estimates the parameters of interest for the study, capturing differences in outcomes for providers p that receive Preschool for All funding in year t_i^* on outcomes k

years later (or k years before, if $k < 0$). Differences are measured relative to year $k = -2$, which corresponds to 2 years prior to the start of Preschool for All funding and the time of the match. I include fixed effects ϕ_k, which are event-time dummies (excluding $k = -2$) to capture common trends in the outcome over time among *both* treatment and comparison ECE providers, and α_g, which includes dummies for matched groups g, such that treatment providers are compared to the 10 most observably similar matched comparison cases. There error term is ε_{igt}.

If comparison cases provide a suitable estimate of the counterfactual for Preschool for All "treated" providers, I expect β_{-4} β_{-3}, which measure differences in the time before the match and before public funding begins, to be indistinguishable from zero. Interaction terms after the match, β_{-1} through β_{+4} shed light on differences between the treatment and comparison groups that arise possibly in response to Preschool for All funding. Recall, many providers begin to report on state pre-K funded services the year prior to the official start of funding. Therefore, I expect outcomes to diverge starting in $k = -1$.

Because I require treatment and comparison providers match in the same calendar year, event time and calendar-year time are aligned within matched groups. This ensures, common time trend estimates come only from comparison providers and not also from already treated or yet to be treated cases. This strategy avoids some of the bias in estimates that can arise from traditional two-way fixed effects models with staggered treatment timing. For an overview, see Baker et al. (2021).

4.8 Results

4.8.1 Is Public Pre-K Offered by New or Preexisting Providers?

To begin, I present descriptive statistics on the 227 ECE providers that received Preschool for All funding from FY2008 to FY2010. Table 4.1 shows the distribution of new vs. preexisting programs by age, the share of programs that are licensed or exempt, and the share that offer pre-K services in a center- or home-based setting. Recall, some public pre-K funded sites are missing from the longitudinal state database on child care providers. I label these providers as "uncategorized." They constitute 16% of Preschool for All funded sites.

As shown in table 4.1, very few Preschool for All providers are new to the ECE sector. Roughly 2% open the year they first receive public pre-K funding, and another 7% entered the ECE sector 1 to 2 years before receiving the grant. In contrast, at least three quarters of Preschool for All providers were already established according to longitudinal state child care records and in business more than 2 years before offering public pre-K. A remaining 16% of public pre-K providers are missing from the state database and may or may not be new to ECE service provision. Even if all uncategorized sites were new to the ECE sector, the overall descriptive pattern suggests a large share of public pre-K services take place in already established child care programs.

Table 4.1 also shows Preschool for All sites located in a mix of licensed and license-exempt settings. License-exempt providers were most commonly public schools. The names of uncategorized programs suggest the majority are license-exempt. Even accounting for this, more than half of public pre-K funded providers (57%) are licensed, suggesting they represent more traditional child care settings. In addition, I find nearly three quarters of public pre-K funded providers are child care centers, as opposed to home-based or uncategorized child

care settings.

Table 4.1: Overview of ECE Providers with Public Pre-K Funding, awarded FY2008-FY2010

	% of Preschool for All
Age of Program	
Newly opened (year of grant)	2.20
Recently opened ($\leqslant$ 2y before grant)	7.05
Established (opened > 2y before grant)	74.89
Uncategorized (missing data)	*15.86*
Program Type	
Licensed	57.71
License-exempt	26.43
Uncategorized (missing data)	*15.86*
Program Setting	
Center-based	74.45
Family child care home	9.69
Uncategorized (missing data)	*15.86*
Observations	227

Notes: Table presents descriptive statistics on child care providers with public pre-K funding through grants awarded under the Preschool for All initiative from FY2008 to FY2010. Newly opened sites are defined as those that appear for the 1st time in the state database during the 1st year of grant funding; recently opened providers appear only 1 or 2 years before; and established providers appear more than 2 years before. Of 227 Preschool for All sites, 36 (16%) do not match to the state registry of child care providers. These are listed as "uncategorized (missing data)."

4.8.2 Does Public Pre-K Funding Receipt Change Preexisting ECE Providers' Operations?

Next, I focus on the sample of established ECE providers that receive public pre-K funding. To begin, I explore characteristics of these providers relative to other ECE providers that do not go on to receive public pre-K funds. Table 4.2 shows this comparison. Column (1) and (2) present averages 2 years before (t=-2) and one year before (t=-1) the providers receive

public pre-K funding. As a benchmark, column (3) presents averages of the same indicators measured among all other providers without public pre-K funding, measured in FY2006. This time point for nonfunded providers captures their characteristics *before* Preschool for All goes into effect (in FY2007). This allows me to compare *prepolicy* differences between the funded and nonfunded groups of providers. Because the large majority of Preschool for All providers are child care centers, I focus only on child care centers in Table 4.2. Table 4.A.1 presents the same comparison among family child care homes.

Perhaps unsurprisingly, differences exist, even at "baseline," between child care centers that later receive public pre-K funding and those that do not. Public pre-K funded providers are less likely to be license-exempt[12] and more likely to report receiving other public funding streams, including state pre-K dollars (from any program, not just Preschool for All), child care subsidies, and Head Start funds. On average, they are more likely to offer care for infants and toddlers, though their propensity to have care available for school-aged children is about the same. In terms of capacity, child care centers that ultimately receive public pre-K funds are larger than those that do not. Average total enrollment capacity for all ages of children as well as the number of slots available per session for preschool-aged children and infants and toddlers are all substantially higher. Last, public pre-K funded centers have more distinct child care sessions at baseline, particularly for preschool-aged children, which perhaps reflects the nature of Preschool for All funding, which covers only part-day care.

Note, provider reports of funding in the INCCRRA registry are less reliable than official State Board of Education records on Preschool for All awards. As such, the two do not always align.[13] However, table 4.2 shows that the share of public pre-K funded providers that report

[12]This difference could result from the fact that 16% of public pre-K funded providers are missing from my matched data. Missing providers were disproportionately more likely to be license-exempt and thus might raise the share of license-exempt providers if included.

[13]For this reason, I rely on State Board of Education records to identify Preschool for All recipients for

Table 4.2: Center-Based Child Care Provider Characteristics by Preschool for All Funding Receipt, Measured Prepolicy

	Public Pre-K Funded		All Other (Not Pre-K Funded)
	$t = -2$	$t = -1$	FY2006
Provider Type (%)			
License-exempt	25.64	29.19	40.54
Accepts Any Funding From (%)			
State pre-K (self-report)	26.92	46.58	14.82
Child care subsidy program	71.15	69.57	60.63
Head Start	16.03	16.15	7.85
Care Available For (%)			
Infants or toddlers	64.74	62.73	42.15
School-aged children	69.87	65.84	64.50
Provider Average (#)			
Total capacity (all ages)	93.08	96.97	72.57
Capacity – preschool-aged	44.45	49.67	33.83
Capacity – infant/toddler	23.55	23.04	12.98
Num. sessions – preschool-aged	1.39	1.53	1.23
Num. sessions – infant/toddler	0.68	0.65	0.45
Observations	156	161	3,744

Notes: Table presents descriptive statistics among center-based child care providers. The column labeled "$t = -2$" presents average values measured two years before the start of public pre-K funding among providers that received Preschool For All funding for the first time between FY2008 and FY2010. The column labeled "$t = -1$" presents average values measured one year before the start of public pre-K funding among the same group. The column labeled "FY2006" presents average values among all other non-pre-K-funded child care centers, measured as of FY2006 (before any Preschool For All funding went into effect in the ECE sector).

receiving state pre-K funds jumps from 27% at $t = -2$ to 47% at $t = -1$, nearly doubling. I take this as evidence that providers may begin reporting on Preschool for All funding and Preschool for All funded services in the INCCRRA data the year before their grant goes into effect; this aligns with the timing of Preschool for All funding announcements, which occur in June before the fiscal year begins, and the timing of INCCRRA data collection, which

analysis purposes.

takes place in June to capture information on the prior school year of completed services. As a result, I consider the time point 2 years before funding begins ($t = -2$) the most accurate representation of "baseline" prepolicy characteristics.

Table 4.3 also presents "baseline" characteristics of neighborhoods where Preschool for All funded providers are located. Estimates come from the 2000 Decennial Census, the earliest year of data before Preschool for All legislation is passed, as well as from the ACS, which offers 5-year estimates between 2005–2009, capturing some of the period before Preschool for All goes into effect. Neighborhood characteristics are measured at the census-tract level.

Table 4.3: Neighborhood Characteristics, by Preschool for All Funding as of 2008–2010

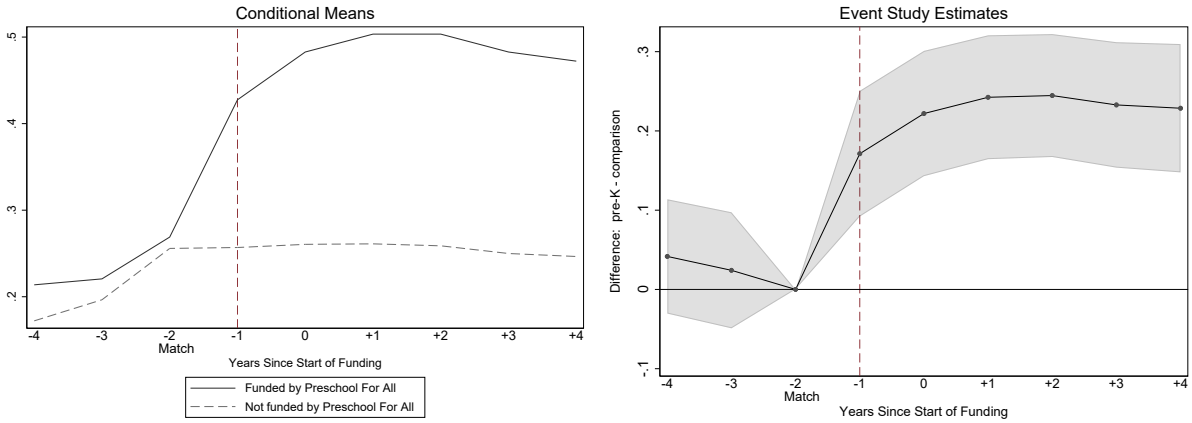
	With Public Pre-K		Without	
	2000 Census	'05–09 ACS	2000 Census	'05–09 ACS
Total population	4,720	4,963	4,180	4,310
Percent under age 5	7.22	7.25	7.06	6.87
Percent identifying as white	73.27	73.45	63.30	66.94
Percent identifying as Black	14.99	16.85	20.67	20.97
Percent identifying as Hispanic	8.69	11.49	12.07	14.04
Percent living in rural area	15.28		14.54	
Median Income		49,180		53,602
Percent below poverty		15.17		15.37
Percent with a BA deg. or more		14.16		17.41
Observations	152	152	2,495	2,495

Notes: Table presents descriptive statistics on Illinois' census tracts with at least one Preschool for All funded provider and tracts without any Preschool For All funded providers. Data come from the 2000 Decennial Census, measured before Preschool For All implementation, and from 2005-2009 American Community Survey (ACS) five-year estimates, which overlaps with implementation. Census tracts in Chicago are excluded because data on Preschool For All funding does not include providers in Chicago. More than one Preschool For All funded provider can locate within a census tract, leading to a lower number of census tracts with public pre-K funding ($n = 152$) than the total number of public pre-K funded ECE providers ($n = 191$).

I find that neighborhoods where a provider receives Preschool for All funding are more populous, have more young children under age 5, have more white residents and fewer Black

or Hispanic residents, and have more residents living in rural areas. In keeping with the targeted nature of Preschool for All funding, neighborhoods with funded providers have lower median family income and lower levels of education (fewer individuals with college-level education or higher). The share of individuals in poverty, however, appears roughly the same in neighborhoods with and without Preschool for All.

Figure 4.1: Propensity Score Matching Estimates: Provider-Reported State Pre-K Funding



Notes: Displays differences in outcomes between Preschool for All and matched comparison providers each year before and after public pre-K funding goes into effect at $t = 0$. The outcome is coded as 1 if providers report to INCCRRA that they receive any form of state pre-K funding, measured in June of each year. Matching occurs at $t = -2$, the time point 2 years before Preschool for All funding is awarded. Public pre-K funded providers are matched to up to 10 observably similar nonfunded providers, based on characteristics shown in Table 4.A under the column "Used to Match: All." Comparison providers are restricted to those in census tracts where no provider received Preschool for All funding. Standard errors from event study estimates are clustered at the provider level and included as shaded areas representing a 95% confidence interval. The vertical dotted line at $t = -1$ reflects the start of possible impacts.

To account for observable differences between funded and nonfunded providers, I turn to my propensity score matching approach. This strategy uses provider- and neighborhood-level variables to identify nonfunded comparison providers most similar to funded providers, but for the receipt of public pre-K funding. Before turning to outcomes of interest, I compare patterns in state pre-K funding reported by providers in the INCCRRA state database. I

compare reports of public pre-K funding between (a) providers identified from State Board of Education records as having received Preschool for All funding in FY2008 to FY2010, and (b) matched nonfunded comparison providers. Figure 4.1 shows mean values conditional on each propensity-score-matched group as well as event study estimates from Equation 4.1. Three key takeaways emerge. First, there is a sharp increase in provider-reported state pre-K funding among Preschool for All providers, suggesting that State Board of Education Records effectively identify providers that receive new public funds. Second, averages between the funded and matched comparison groups appear similar at $t = -3$ and $t = -4$, 3 and 4 years before funding went into place, time points that were not matched. This suggests the match at $t = -2$ does an effective job identifying providers similar in other time points prior to Preschool for All implementation. Third, Figure 4.1 shows that providers begin reporting changes in state pre-K funding the year before it takes effect. This is consistent with descriptive patterns seen earlier and confirms changes in response to public pre-K funding are likely to be reported at $t = -1$, the year prior to funds taking effect.

Propensity Score Matching Estimates: Propensity to Remain Open

With these patterns in mind, I turn to the results as to whether public pre-K funding changes providers' propensity to remain in business. Figure 4.2 presents conditional means and event study estimates. At $t = -2$, the time of the match, 100% of Preschool for All providers and matched comparison providers are, by definition, in business. Patterns in terms of propensity to be in business are nearly identical across the two groups leading up to the match. However, patterns diverge beginning at $t = -1$ and continuing through $t = 4$, such that 4 years after public pre-K funding goes into effect, providers with the funding are 15 percentage points more likely to remain open than providers in the matched comparison group. Based on

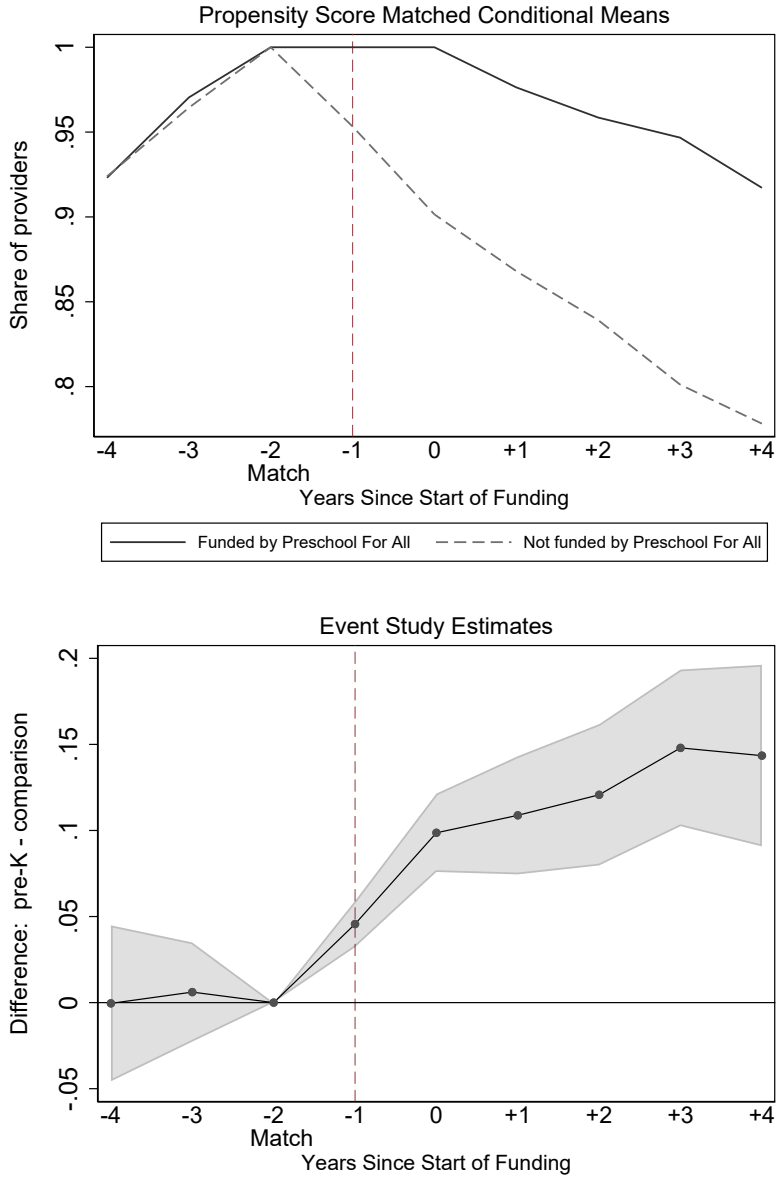
these findings alone, Preschool for All funding expands access to ECE services by virtue of preventing ECE programs that would have otherwise closed from doing so.

Evidence of differential attrition between Preschool for All providers and their matched comparison cases creates measurement challenges for studying other program-level outcomes. In other words, if I were to compare these same two groups over time, the composition of the comparison group would change in nonrandom ways, relative to the group of Preschool for All providers. Then, results would partially capture changes in outcomes due to differential selection out of the sample between Preschool for All providers and the matched comparison group. To partially control for this, I limit subsequent analyses to a balanced panel of providers, both Preschool for All and matched comparison providers that I observe for the entire study window. In this way, I ensure that changes in the outcomes do not result from changes in the composition of the treatment and comparison groups. Two downsides to this approach are that (1) it limits my sample size and reduces the number of possible comparison providers to which Preschool for All providers can match, and (2) providers in the comparison group that remain in the business may be relatively "stronger," or more financially robust given that these are the ones that avoid closure. Both of these facts mean that I am less likely to detect changes in outcomes, making my estimates more conservative.

Provider Capacity

Among the balanced panel of ECE providers, I next explore whether public pre-K funding increases providers' enrollment capacity relative to matched comparison providers without increased funding. Figure 4.3 presents results. Panel A focuses on providers' total capacity, or the maximum number of children they can have in care at any one time. Panel B shows preschool-aged capacity, which reflects the maximum number of preschoolers allowed to enroll

Figure 4.2: Provider Remains in Business (0/1)



Notes: Displays differences in outcomes between Preschool for All and matched comparison providers each year before and after public pre-K funding goes into effect at $t = 0$. The outcome is coded as 1 if providers appear in the INCCRRA database and 0 if they do not. Matching occurs at $t = -2$, the time point 2 years before Preschool for All funding is awarded. Public pre-K funded providers are matched to up to 10 observably similar nonfunded providers based on characteristics show in Table 4.A under the column "Used to Match: All." Comparison providers are restricted to those in census tracts where no provider received Preschool for All funding. Standard errors from event study estimates are clustered at the provider level and included as shaded area representing a 95% confidence interval. The vertical dotted line at $t = -1$ reflects the start of possible impacts.

in a single child care session. Average capacity among public pre-K funded and matched nonfunded providers appears similar over time. It does not diverge after pre-K funding goes into effect. As such, event study estimates show no detectable change in overall or preschool-specific capacity among providers that receive Preschool for All funding.

Other ECE Services

Next, I explore whether providers shift other aspects of their ECE services or operations. Outcomes include the number of distinct child care sessions they offer, the types of other public funding sources they receive, or availability to care for other age groups of children. Across these outcomes, public pre-K funding appears to increase the number of distinct child care sessions available for preschool-aged children but has no detectable influence on any other area of service provision.

Figure 4.4 shows differences in the number of distinct preschool-aged child care sessions offered by public pre-K funded providers relative to matched comparison providers. This measure captures differences between a single full-day session, for example, and 2 part-day sessions within a single day. Growth in number of child care sessions among public pre-K providers tracks funding requirements closely; Preschool for All funds cover 2.5 hours of pre-K services for children, suggesting part-day sessions would increase if they are not already in place. Notably, children can enroll in back-to-back part-day sessions, so a shift to more distinct sessions does not necessarily imply less total hours in care per child. However, with an increase in the distinct number of sessions, families can opt to receive care during only one of the sessions. As a result, an increase in sessions may suggest an increase in the odds that children spend less time in care but that providers manage to serve more children on average across the school day.

As shown in Figure 4.5, patterns in ECE availability for infants and toddlers do not appear to change in response to public pre-K funding. I do not detect differences in whether a provider offers *any* care for infants/toddlers, more or less capacity for infants/toddlers per session, or greater or fewer numbers of distinct child care sessions. Similarly, providers with public pre-K funding do not appear reliably more or less likely to offer care for school-aged children, though estimates are noisy (see Figure 4.6). Last, I turn to whether public pre-K funding receipt changes providers' likelihood of reporting they receive funding from Head Start or accept government-funded payments through the child care subsidy program, shown in Figure 4.7. Again, estimates are noisy, and I am unable to distinguish any meaningful change in other sources of public funding in response to Preschool for All receipt.

4.9 Conclusion

In light of rapid expansion in state-level public pre-K programs, this paper explores the implementation of Illinois' Preschool for All initiative designed to serve children living in families with low incomes and other socioeconomic risk factors. I leverage data at the ECE provider level, connecting records from the Illinois State Department of Education to a longitudinal registry of the universe of child care providers in Illinois. First, I show that the majority (at least 75%) of public pre-K services funded under the Preschool for All initiative take place in already established ECE provider settings.

Second, I explore direct effects of new public pre-K funding among providers that existed prior to pre-K expansion. I use a propensity score matching approach that assigns public pre-K funded providers to observably similar nonfunded providers and find sizable increases in the likelihood that public pre-K funded providers remain in the ECE sector. Four years after Preschool for All funding receipt, more than 90% of Preschool for All funded providers

continue to operate, while fewer than 80% of matched comparison providers remain in business. However, among a balanced sample of providers that continue to remain in business, I find no detectable change in Preschool for All providers' total enrollment capacity, their likelihood of caring for children in other age groups (infants/toddlers or school-aged children), or their likelihood of receiving other sources of public ECE funds. The main change Preschool for All providers make in response to the funding is to increase the number of distinct child care sessions offered for preschool-aged children within the school day. This change is closely related to Preschool for All as a funding source, which covers half-day care for eligible children and thus may induce providers to offer more sessions that are shorter in length.

While public pre-K expansion is often predicated on the notion that newly available funding will create new programs and expand ECE sector capacity, the current study's findings suggest a more nuanced policy response in Illinois. Notably, public pre-K expansion largely funded existing providers. Yet, among these providers, it helped families maintain access to care by preventing closures. Importantly, this finding occurs in the context of the Great Recession, when job and income loss may have reduced families' ability to pay for child care out of pocket. Providers that did not receive public pre-K funding starting in FY2008-FY2010 may have struggled to stay in business in the face of declining demand among private-pay customers during the economic downturn. As such, I would expect smaller impacts on provider attrition during other, more economically robust periods of time. Nonetheless, the finding that public pre-K funding helped stabilize providers demonstrates the importance of public investments in ECE infrastructure over the business cycle.

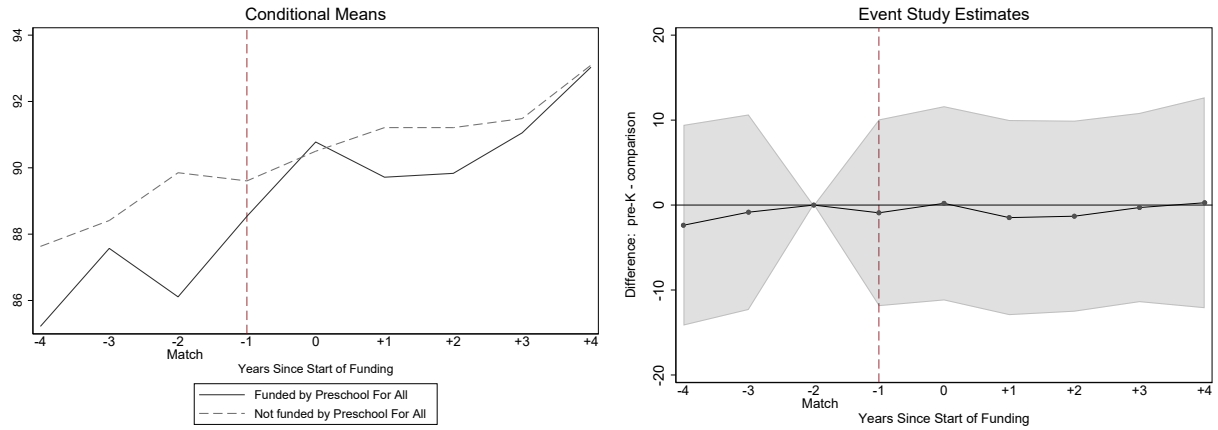
In addition, study findings suggest Preschool for All may have expanded the total number of children enrolled in ECE by offering more distinct child care sessions but for fewer hours

per day. This finding raises important questions about the goals of public pre-K. For example, recent experimental evidence shows full-day pre-K improves children's school readiness skills to a greater degree than half-day care (Atteberry et al., 2019). In addition, full-day care tends to better align with working parents' child care needs. However, in the face of limited resources, do the benefits of offering more children a shorter amount of time in pre-K outweigh the costs of offering fewer children a full day of care? To begin to answer this question, future research should explore not just whether public pre-K programs increase the number of slots funded providers offer, nor simply the number of sessions, but rather total hours of care children receive in the setting.

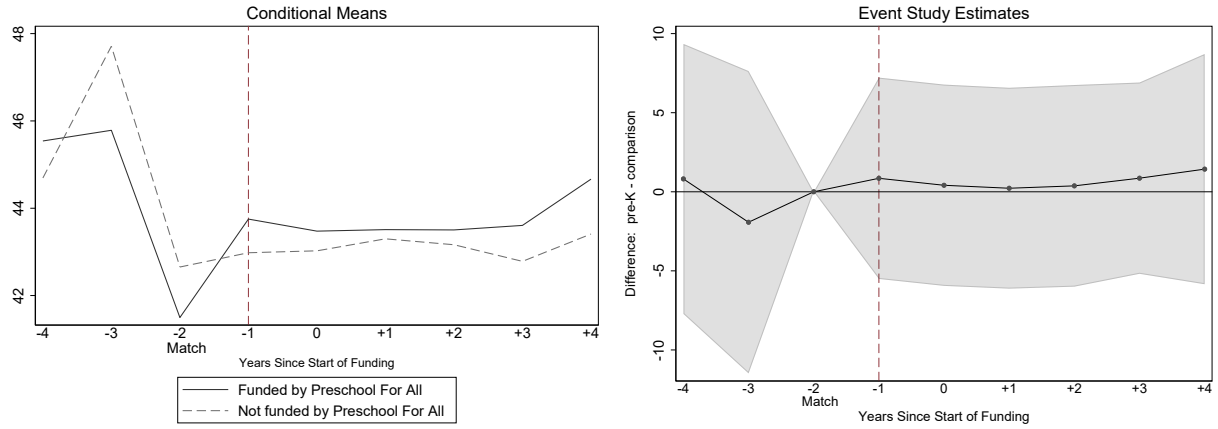
Finally, in contrast to prior research, which investigates aggregate impacts of public pre-K programs on the ECE sector at large, the current study isolates the direct influence of public pre-K funding on funded providers. As such, the findings begin to inform our understanding of how and why the ECE sector as a whole, including providers that do not receive funding, might respond to state pre-K expansion. Based on findings that Preschool for All providers are more likely to remain in business, there is clearly scope for competitive spillover effects on nonfunded providers in the same local markets. These providers, faced with competition from a more stable provider, may close at even higher rates than the comparison providers in this study (those where no Preschool for All program had located nearby). Note, while I exclude nonfunded providers in the census tracts with Preschool for All from this analysis, future research should explore whether Preschool for All funding influences nearby nonfunded providers through competitive effects. A better understanding of the direct and competitive effects of public pre-K expansion should enable policymakers and practitioners to more carefully deploy policy strategies to increase families' access to early care and education.

Figure 4.3: Propensity Score Matching Estimates: Maximum Capacity to Enroll Children

(a) Total Capacity (max. # children at any given time)

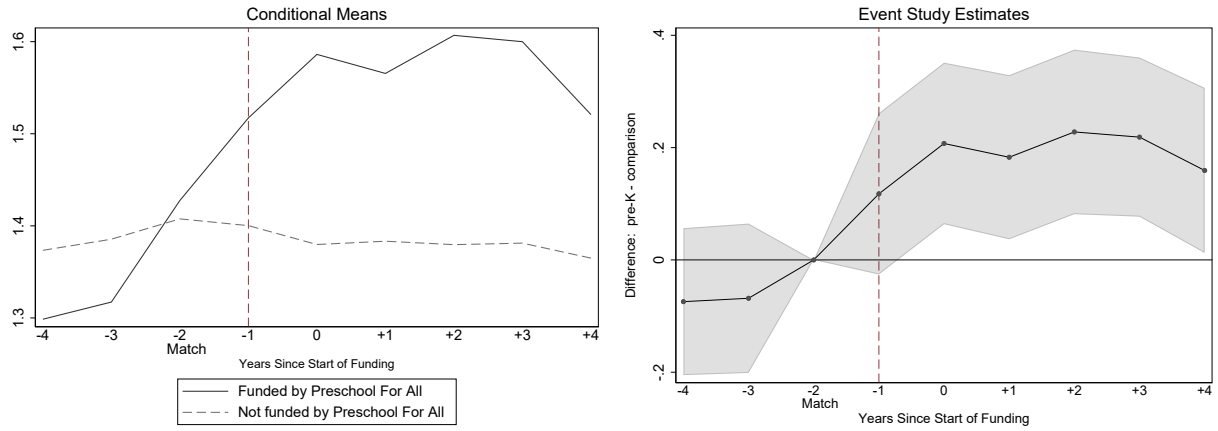


(b) Preschool-Aged Capacity (max. # preschoolers per session)



Notes: Displays differences in outcomes between Preschool for All and matched comparison providers each year before and after public pre-K funding goes into effect at $t = 0$ among a balanced sample of providers. Outcomes include: (1) total provider capacity, which reflects the number of children who can be in care at any given time; and (2) preschool-aged capacity, which reflects the maximum number of children ages 3 to 5 who can be in a single child care session. Matching occurs at $t = -2$, the time point 2 years before Preschool for All funding is awarded. Public pre-K funded providers are matched to up to 10 observably similar nonfunded providers, based on characteristics shown in Table 4.A under the column "Used to Match: Balanced." Comparison providers are restricted to those in census tracts where no provider received Preschool for All funding. Standard errors from event study estimates are clustered at the provider level and included as shaded area representing a 95% confidence interval. The vertical dotted line at $t = -1$ reflects the start of possible impacts.

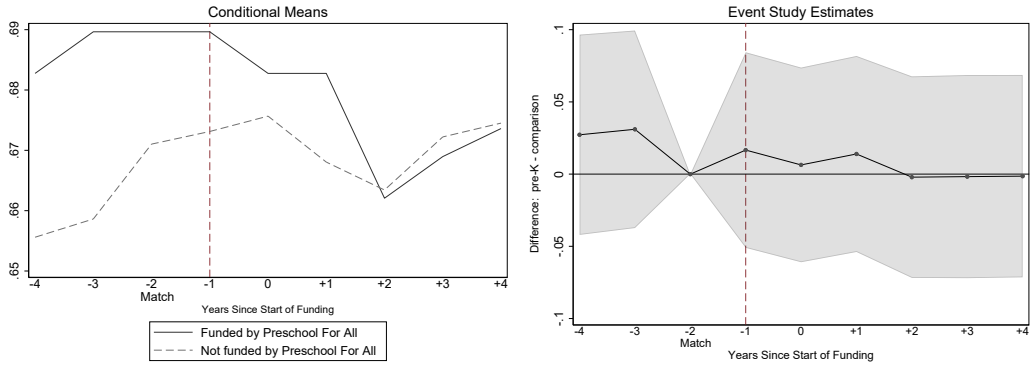
Figure 4.4: Propensity Score Matching Estimates: Number of Child Care Sessions for Preschool-Aged Children



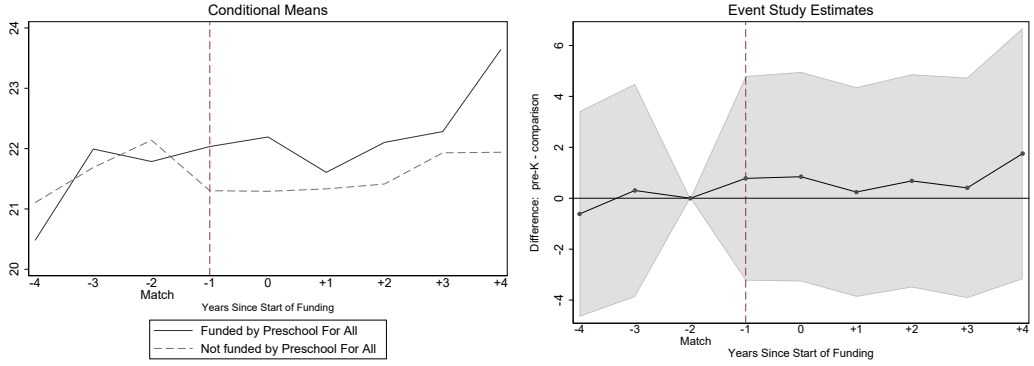
Notes: Displays differences in outcomes between Preschool for All and matched comparison providers each year before and after public pre-K funding goes into effect at $t = 0$ among a balanced sample of providers. The outcome captures the number of distinct child care sessions a provider offers for preschool-aged children. For additional notes, see Figure 4.3.

Figure 4.5: Propensity Score Matching Estimates: ECE Services for Infants/Toddlers

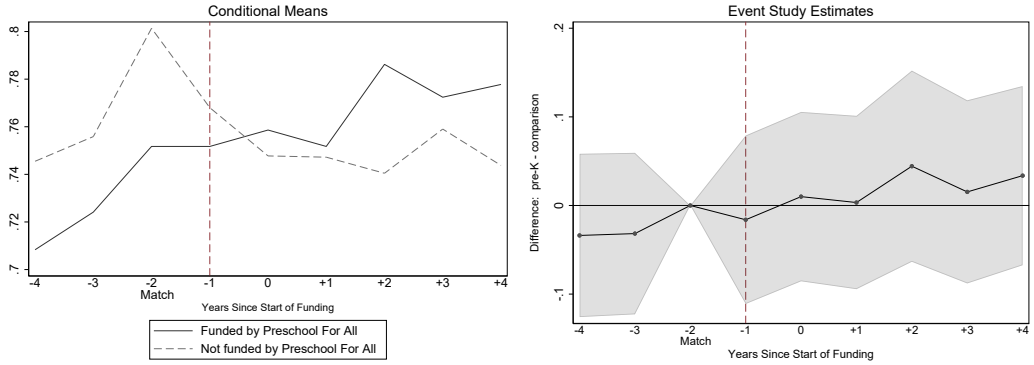
(a) Any Care Available for Infants/Toddlers



(b) Infant/Toddler Capacity (max. per session)

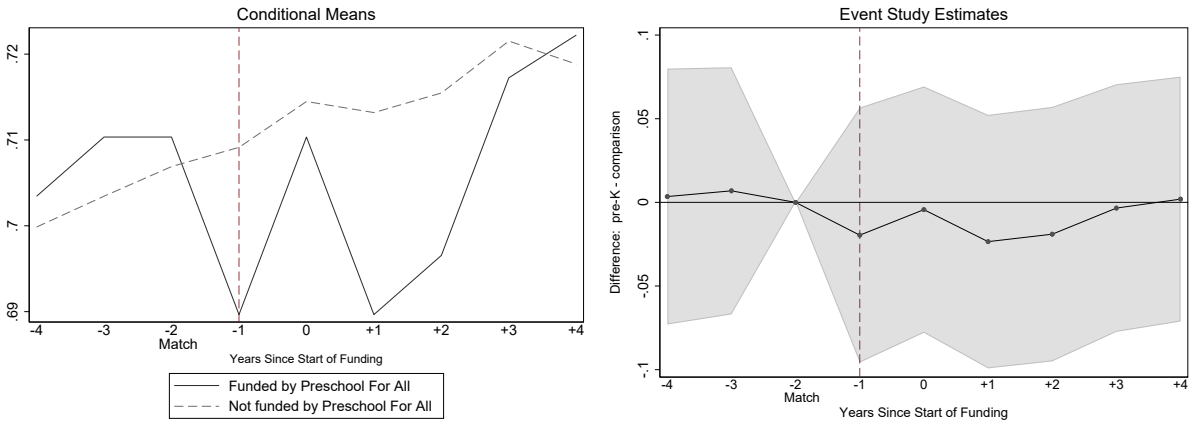


(c) Number of Child Care Sessions for Infants/Toddlers



Notes: Displays differences in outcomes between Preschool for All and matched comparison providers before and after public pre-K funding goes into effect ($t = 0$) among a balanced sample of providers. Outcomes include: (1) whether a provider offers any care for infants/toddlers; (2) the maximum number of children ages 6 weeks to <3 years old who can be in a child care session (infant/toddler capacity); and (3) the number of distinct child care sessions a provider offers for infants/toddlers. For additional notes, see Figure 4.3.

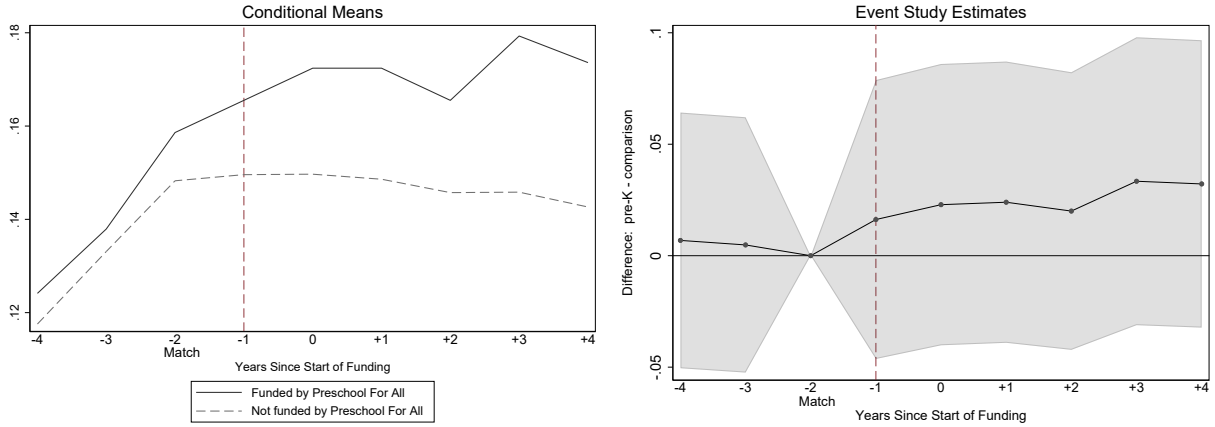
Figure 4.6: Propensity Score Matching Estimates: Any Care Available for School-Aged Children



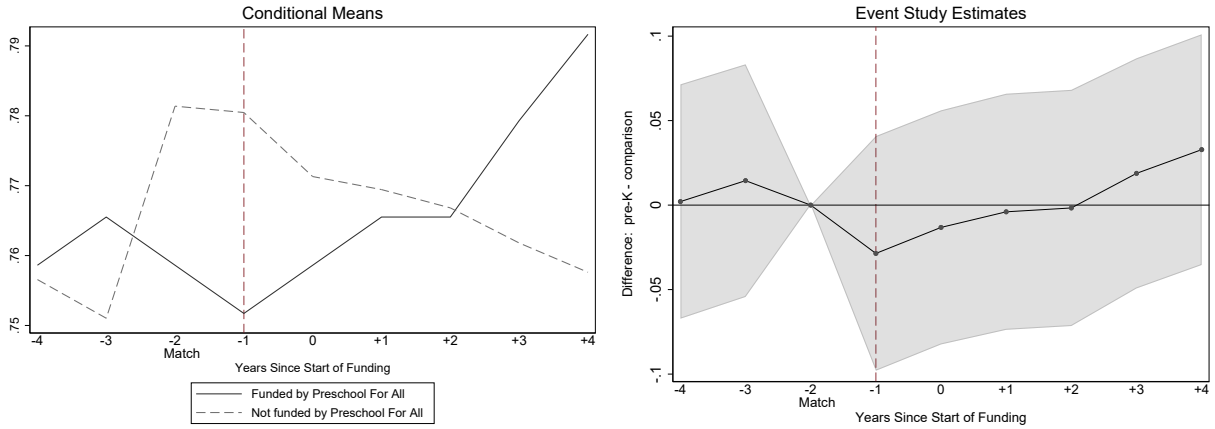
Notes: Displays differences in outcomes between Preschool for All and matched comparison providers each year before and after public pre-K funding goes into effect at $t = 0$ among a balanced sample of providers. The outcome is an indicator equal to 1 if the provider offers any care for school-aged children. For additional notes, see Figure 4.3.

Figure 4.7: Propensity Score Matching Estimates: Other Public Funding Sources

(a) Any Head Start Funding



(b) Accepts Child Care Subsidies



Notes: Displays differences in outcomes between Preschool for All and matched comparison providers each year before and after public pre-K funding goes into effect at $t = 0$ among a balanced sample of providers. Outcomes include: (1) an indicator equal to 1 if the provider reports funding from the Head Start program; and (2) an indicator equal to 1 if the provider reports they accept government-funded child care subsidies. For additional notes, see Figure 4.3.

Appendix

4.A Supplemental Tables

Table 4.A.1: Family Child Care Provider Characteristics by Preschool for All Funding Receipt, Measured Prepolicy

	Public Pre-K Funded		All Other (Not Pre-K Funded)
	$t = -2$	$t = -1$	FY2006
Accepts Any Funding From (%)			
State pre-K (self-report)	0.00	0.00	0.01
Child care subsidy program	90.00	86.36	79.15
Head Start	0.00	0.00	0.02
Care Available For (%)			
Infants or toddlers	90.00	90.91	98.30
School-aged children	90.00	86.36	95.41
Provider Average (#)			
Total capacity (all ages)	11.20	11.18	8.36
Enrollment – preschool-aged	2.80	3.18	2.07
Enrollment – infant/toddler	3.25	3.14	2.49
Num. sessions – preschool-aged	1.30	1.09	1.17
Num. sessions – infant/toddler	1.20	1.14	1.18
Observations	20	22	8,412

Notes: Table presents descriptive statistics among family child care providers. The column labeled $t = -2$ presents average values measured two years before the start of public pre-K funding among providers that received Preschool For All funding for the first time between FY2008 and FY2010. The column labeled $t = -1$ presents average values measured one year before the start of public pre-K funding among the same group. The column labeled FY2006 presents average values among all other non pre-K funded family child care providers, measured as of FY2006 (before Preschool For All expansion).

Table 4.A.2: Variables Used to Predict Propensity Scores Among the Overall and Balanced Samples of ECE Providers

Variables	Notes	Used to Match: All	Used to Match: Balanced
Any Head Start funding		x	x
Any state pre-K funding		x	x
Accepts child care subsidies		x	x
Accepts vouchers from child protective services	Refers to vouchers from the Department of Child and Family Services in Illinois	x	
License-exempt*	Operates in a public school, college, religious institution or setting otherwise exempt from required ECE licensing	x	x
Serves children under age 3		x	x
Serves school-aged children		x	x
Total capacity	Maximum number of all children a provider can care for at any one time	x	x
Preschool-aged capacity	Maximum number of preschool-aged children a provider can care for at any one time. Reported only among center-based providers.	x	x
Preschool-aged capacity in year prior	Same as above, measured in the year before the year of the match.		x
Infant/toddler capacity	Maximum number of children under age 3 a provider can care for at any one time. Reported only among center-based providers.	x	x
Preschool-aged enrollment	Number of enrolled preschool-aged children. Measured only among family child care providers.	x	
Infant/toddler enrollment	Number of enrolled infants and toddlers. Measured only among family child care providers.	x	

Table 4.A.2: Variables Used to Predict Propensity Scores Among the Overall and Balanced Samples *(Continued)*

Variables	Notes	Used to Match: All	Balanced
Any part-time only care	Measured separately for infants/ toddlers and preschool-aged children.	x	
Any full-time only care	Measured separately for infants/ toddlers and preschool-aged children.	x	
Both part- and full-time care	Measured separately for infants/ toddlers and preschool-aged children.	x	
Number of sessions*	Counts distinct sessions (i.e., multiple part-time sessions, or combinations of full- and part-time sessions). Measured separately for infants/toddlers and preschool-aged children.	x	x
1st year provider appears in the data		x	
Provider county location		x	x
Census tract total population	Data from 2000 Census	x	x
Percent under age 5 in census tract	Data from 2000 Census	x	x
Percent identifying as white in census tract	Data from 2000 Census	x	
Percent identifying as Black in census tract	Data from 2000 Census	x	x
Percent identifying as Hispanic in census tract	Data from 2000 Census	x	x
Percent living in a rural area in census tract*	Data from 2000 Census	x	
Median income in census tract*	Data from 2005-2009 ACS	x	x
Percent with a BA degree or higher in census tract*	Data from 2005-2009 ACS	x	x

Table 4.A.2: Variables Used to Predict Propensity Scores Among the Overall and Balanced Samples *(Continued)*

Variables	Notes	Used to Match: All	Balanced
Percent living below poverty line in census tract*	Data from 2005-2009 ACS	x	x
→ interaction terms for providers willing to accept child care subsidies and variables marked with a * above			x

Notes: Lists variables used to calculate the propensity that an ECE provider received Preschool For All funding in a given year. The column "All" shows variables marked with an x if used in calculating propensity scores for the purposes of matching the overall sample of Preschool For All providers to nonfunded providers, regardless of attrition. The column "Balanced" shows variables marked with an x if used in calculating propensity scores for the purposes of matching among the balanced sample where providers were required to be observed 4 years before and 4 years after the match.

Bibliography

Agüero, J. M. and Marks, M. S. (2011). Motherhood and Female Labor Supply in the Developing World: Evidence from Infertility Shocks. *The Journal of Human Resources*, 46(4):800–826.

Andresen, M. E. and Nix, E. (2020). What Causes the Child Penalty? Evidence from Same Sex Couples and Policy Reforms. page 69.

Angelov, N., Johansson, P., and Lindahl, E. (2016). Parenthood and the Gender Gap in Pay. *Journal of Labor Economics*, 34(3):545–579.

Angrist, J. D. and Evans, W. N. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *The American Economic Review*, 88(3):450–477.

Antecol, H., Bedard, K., and Stearns, J. (2018). Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? *American Economic Review*, 108(9):2420–41.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., and Schady, N. (2016). Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics*, 131(3):1415–1453.

Atteberry, A., Bassok, D., and Wong, V. C. (2019). The effects of full-day prekindergarten: Experimental evidence of impacts on children's school readiness. *Educational Evaluation and Policy Analysis*, 41(4):537–562.

Azmat, G. and Ferrer, R. (2017). Gender gaps in performance: Evidence from young lawyers. *Journal of Political Economy*, 125(5):1306–1355.

Bacolod, M., Heissel, J. A., and Sullivan, R. (2020). Mothers in the Military: The Effect of Department of Defense Maternity Policy on Leave Uptake and Substitution. Manuscript submitted for publication.

Bailey, M., Byker, T., Patel, E., and Ramnath, S. (2019). The Long-Term Effects of California's 2004 Paid Family Leave Act on Women's Careers: Evidence from U.S. Tax Data. Technical Report w26416, National Bureau of Economic Research, Cambridge, MA.

Baker, A., Larcker, D. F., and Wang, C. C. (2021). How much should we trust staggered difference-in-differences estimates? *Available at SSRN 3794018*.

Balser, C., Hall, C., and Bukowinski, A. (2020). The Effect of Expanding Paid Maternity Leave on Maternal Health: Evidence From the United States Air Force & Army. SSRN Scholarly Paper ID 3601018, Social Science Research Network, Rochester, NY.

Barnett, W. S., Hustedt, J. T., Friedman, A. H., Stevenson Boyd, J., and Ainsworth, P. (20007). The State of Preschool 2007: State Preschool Yearbook. *National Institute for Early Education Research*.

Barth, E., Kerr, S. P., and Olivetti, C. (2017). The Dynamics of Gender Earnings Differentials: Evidence from Establishment Data. Technical Report w23381, National Bureau of Economic Research, Cambridge, MA.

Bassok, D. (2012). Competition or Collaboration? Head Start Enrollment During the Rapid Expansion of State Pre-Kindergarten. *Educational Policy*, 26(1):96–116.

Bassok, D., Dee, T. S., and Latham, S. (2019). The Effects of Accountability Incentives in Early Childhood Education. *Journal of Policy Analysis and Management*, 0(0):1–29.

Bassok, D., Fitzpatrick, M., and Loeb, S. (2014). Does State Preschool Crowd-Out Private Provision? The impact of Universal Preschool on the Childcare Sector in Oklahoma and Georgia. *Journal of Urban Economics*, 83:18–33.

Bassok, D., Miller, L. C., and Galdo, E. (2016). The Effects of Universal State Pre-Kindergarten on the Child Care Sector: The Case of Florida's Voluntary Pre-Kindergarten Program. *Economics of Education Review*, 53:87–98.

Baum, C. L. and Ruhm, C. J. (2016). The effects of paid family leave in california on labor market outcomes. *Journal of Policy Analysis and Management*, 35(2):333–356.

Bertrand, M., Goldin, C., and Katz, L. F. (2010). Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors. *American Economic Journal: Applied Economics*, 2(3):228–255.

Bohn, C. M., Roehrig, A. D., and Pressley, M. (2004). The First Days of School in the Classrooms of Two More Effective and Four Less Effective Primary-Grades Teachers. *The Elementary School Journal*, 104(4):269–287.

Briggs, D. C. and Alzen, J. L. (2019). Making Inferences About Teacher Observation Scores Over Time. *Educational and Psychological Measurement*.

Bronars, S. G. and Grogger, J. (1994). The Economic Consequences of Unwed Motherhood:

Using Twin Births as a Natural Experiment. *The American Economic Review*, 84(5):1141–1156.

Brown, J. H. (2019). Does Public Pre-K Have Unintended Consequences on the Child Care Market for Infants and Toddlers? *Job Market Paper*.

Buell, M., Han, M., and Vukelich, C. (2017). Factors Affecting Variance in Classroom Assessment Scoring System Scores: Season, Context, and Classroom Composition. *Early Child Development and Care*, 187(11):1635–1648.

Bullinger, L. R. (2019). The effect of paid family leave on infant and parental health in the united states. *Journal of health economics*, 66:101–116.

Bureau of Labor Statistics (2017a). Physical strength required for jobs in different occupations in 2016. Bureau of Labor Statistics (BLS) and United States Department of Labor.

Bureau of Labor Statistics (2017b). Standing or walking versus sitting on the job in 2016. Bureau of Labor Statistics (BLS) and United States Department of Labor.

Bureau of Labor Statistics (2020). Employment characteristics of families summary. Bureau of Labor Statistics (BLS) and United States Department of Labor.

Burgett, L., Doan, K., and Metcalf, J. (2018). Early Childhood Block Grant - Preschool for All: FY19 Open, Competitive Technical Assistance Webinar.

Butikofer, A., Riise, J., and Skira, M. (2018). The impact of paid maternity leave on maternal health. *NHH Dept. of Economics Discussion Paper*, (04).

Byker, T. S. (2016). Paid Parental Leave Laws in the United States: Does Short-Duration

Leave Affect Women's Labor-Force Attachment? *American Economic Review*, 106(5):242–246.

Casabianca, J. M., Lockwood, J. R., and McCaffrey, D. F. (2015). Trends in Classroom Observation Scores. *Educational and Psychological Measurement*, 75(2):311–337.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., and Pianta, R. C. (2013). Effect of Observation Mode on Measures of Secondary Mathematics Teaching. *Educational and Psychological Measurement*, 73(5):757–783.

Cascio, E. U. and Schanzenbach, D. W. (2013). The Impacts of Expanding Access to High-Quality Preschool Education. Technical report, The Brookings Institution, Washington, DC.

Cash, A. H. and Pianta, R. C. (2014). The Role of Scheduling in Observing Teacher–Child Interactions. *School Psychology Review*, 43(4):428–449. Publisher: Routledge _eprint: https://doi.org/10.1080/02796015.2014.12087414.

Castro, S., Granlund, M., and Almqvist, L. (2017). The Relationship Between Classroom Quality-Related Variables and Engagement Levels in Swedish Preschool Classrooms: A Longitudinal Study. *European Early Childhood Education Research Journal*, 25(1):122–135. Publisher: Routledge _eprint: https://doi.org/10.1080/1350293X.2015.1102413.

Cattaneo, M. D., Jansson, M., and Ma, X. (2018). Manipulation testing based on density discontinuity. *The Stata Journal*, 18(1):234–261.

Congressional Research Service (2016). Preschool Development Grants (FY2014-FY2016) and Race to the Top—Early Learning Challenge Grants (FY2011-FY2013). CRS Report R44008.

Cools, S., Markussen, S., and Strøm, M. (2017). Children and Careers: How Family Size Affects Parents' Labor Market Outcomes in the Long Run. *Demography; Silver Spring*, 54(5):1773–1793.

Cruces, G. and Galiani, S. (2007). Fertility and female labor supply in Latin America: New causal evidence. *Labour Economics*, 14(3):565–573.

Cunha, J. M., Shen, Y.-C., and Burke, Z. R. (2018). Contrasting the impacts of combat and humanitarian assistance/disaster relief missions on the mental health of military service members. *Defence and Peace Economics*, 29(1):62–77.

Curby, T. W., Grimm, K. J., and Pianta, R. C. (2010). Stability and Change in Early Childhood Classroom Interactions During the First Two Hours of a Day. *Early Childhood Research Quarterly*, 25(3):373–384.

Curby, T. W., Rudasill, K. M., Edwards, T., and Pérez-Edgar, K. (2011). The Role of Classroom Quality in Ameliorating the Academic and Social Risks Associated with Difficult Temperament. *School Psychology Quarterly*, 26(2):175–188.

Cáceres-Delpiano, J. (2006). The Impacts of Family Size on Investment in Child Quality. *The Journal of Human Resources*, 41(4):738–754.

Department of Defense (2018). 2018 demographics report: Profile of the military community". Technical report, Department of Defense (DoD), Office of the Deputy Assistant Secretary of Defense for Military Community and Family Policy (ODASD (MC&FP)), under contract with ICF (https://www.icf.com/work/human-capital.

Derrick-Mills, T., Burchinal, M., Peters, H., De Marco, A., Forestieri, N., Fyffe, S., and

Woods, T. (2016). Early Implementation of the Head Start Designation Renewal System: Volume i. Technical Report OPRE Report 2016-75a, Urban Institute, Washington, DC.

Dodge, K. (2017). The Current State of Scientific Knowledge on Pre-Kindergarten Effects.

Donovan, S. A. (2019). Paid Family Leave in the United States. Report 44835, Congressional Research Services, Washington, DC.

Downer, J., Sabol, T., and Hamre, B. (2010). Teacher–Child Interactions in the Classroom: Toward a Theory of Within- and Cross-Domain Links to Children's Developmental Outcomes. *Early Education and Development*, 21:699–723.

Dynarski, S., Hyman, J., and Schanzenbach, D. W. (2013). Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion. *Journal of policy Analysis and management*, 32(4):692–717.

Early Childhood Learning & Knowledge Center (2020). A National Overview of Grantee CLASS Scores in 2020. Technical report, Office of Head Start, An Office of the Administration for Children and Families, Washington, DC.

Fitzpatrick, M. D. (2010). Preschoolers Enrolled and Mothers at Work? The Effects of Universal Prekindergarten. *Journal of Labor Economics*, 28(1):51–85.

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G., and Gardiner, B. A. (2021). The State of Preschool 2020: State Preschool Yearbook. *National Institute for Early Education Research.*

Gallen, Y. (2018). Motherhood and the gender productivity gap. *Becker Friedman Institute for Research in Economics Working Paper*, (2018-41).

Gandhi, J., Raver, C. C., Abenavoli, R. M., Morris, P. A., and Meyer, L. M. (2020). Variations in Pre-Kindergarten Classroom Quality Ratings across the School Year: Observation Ratings from New York City's Pre-K for All. *Early Education and Development*, 0(0):1–20. Publisher: Routledge _eprint: https://doi.org/10.1080/10409289.2020.1829291.

Hamilton, B. E., Martin, J. A., Osterman, M. J., and Rossen, L. M. (2019). Births: Provisional Data for 2018. Vital Statistics Rapid Release 007, U.S. Department of Health and Human Services, Washington, DC.

Hamre, B., Hatfield, B., Pianta, R., and Jamil, F. (2014). Evidence for General and Domain-Specific Elements of Teacher–Child Interactions: Associations with Preschool Children's Development. *Child development*, 85(3):1257–1274.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., and Downer, J. T. (2008). Building a Science of Classrooms: Application of the CLASS Framework in over 4,000 U.S. Early Childhood and Elementary Classrooms. *Foundation for Childhood Development*, 30.

Heckman, J. J. (2006). Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*, 312(5782):1900–1902.

Illinois Action for Children (2021). How to Start a Child Care Center.

Illinois State Board of Education (2007). Preschool For All: Nuturing Illinois' Promise. A Primer for Providers.

Jacobsen, J. P., Pearce, J. W., and Rosenbloom, J. L. (1999). The Effects of Childbearing on Married Women's Labor Supply and Earnings: Using Twin Births as a Natural Experiment. *The Journal of Human Resources*, 34(3):449–474.

Jessen-Howard, S., Malik, R., Workman, S., and Hamm, K. (2018). Understanding Infant and Toddler Child Care Deserts. Technical report, Center for American Progress, Washington, D.C.

Jukic, A. M., Baird, D. D., Weinberg, C. R., McConnaughey, D. R., and Wilcox, A. J. (2013). Length of human pregnancy and contributors to its natural variation. *Human reproduction*, 28(10):2848–2855.

Kim, S. and Moser, P. (2020). Women in science: Lessons from the baby boom. National Bureau of Economic Research Summer Institute (NBER SI) Session on Gender in the Economy.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., and Zweimüller, J. (2020). Do family policies reduce gender inequality? evidence from 60 years of policy experimentation. Technical report, National Bureau of Economic Research.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., and Zweimüller, J. (2019a). Child Penalties Across Countries: Evidence and Explanations. NBER Working Paper w25524, National Bureau of Economic Research, Cambridge, MA.

Kleven, H., Landais, C., and Søgaard, J. E. (2019b). Children and Gender Inequality: Evidence from Denmark. *American Economic Journal: Applied Economics*, 11(4):181–209.

Kuhfeld, M. and Soland, J. (2021). The Learning Curve: Revisiting the Assumption of Linear Growth during the School Year. *Journal of Research on Educational Effectiveness*, 14(1):143–171.

Lloyd, D. and Joseph, L. (2014). The Erosion of Early Childhood Investments in Illinois. *Fiscal Policy Center at Voices for Children*, Special Report.

Ludwig, J. and Miller, D. L. (2007). Does Head Start Improve Children's Life Chances? Evidence From a Regression Discontinuity Design. *The Quarterly journal of economics*, 122(1):159–208.

Malik, R., Hamm, K., Schochet, L., Novoa, C., Workman, S., and Jessen-Howard, S. (2018). America's Child Care Deserts in 2018. Technical report, Center for American Progress, Washington, D.C.

Malmberg, L.-E., Hagger, H., Burn, K., Mutton, T., and Colls, H. (2010). Observed classroom Quality During TeacherEducation and Two Years of Professional Practice. *Journal of Educational Psychology*, 102(4):916–932.

Malmberg, L.-E., Hagger, H., and Webster, S. (2014). Teachers' Situation-Specific Mastery Experiences: Teacher, Student Group and Lesson Effects. *European Journal of Psychology of Education*, 29(3):429–451.

Mashburn, A. J. (2017). Evaluating the Validity of Classroom Observations in the Head Start Designation Renewal System. *Educational Psychologist*, 52(1):38–49.

Melchiorre, K., Sharma, R., Khalil, A., and Thilaganathan, B. (2016). Maternal cardiovascular function in normal pregnancy: evidence of maladaptation to chronic volume overload. *Hypertension*, 67(4):754–762.

Memon, H. U. and Handa, V. L. (2013). Vaginal childbirth and pelvic floor disorders. *Women's health*, 9(3):265–277.

Meyer, J. P., Cash, A. H., and Mashburn, A. (2011). Occasions and the Reliability of Classroom Observations: Alternative Conceptualizations and Methods of Analysis. *Educational Assessment*, 16(4):227–243.

National Center on Early Childhood Quality Assurance (2017). History of QRIS Growth Over Time. Technical report, Fairfax, VA.

Office of Head Start (2013). Report to Congress on Head Start Monitoring, Fiscal Year 2013 | ECLKC. Technical report, The Administration for Children and Families, Department of Health and Human Services.

Office of Head Start (2016). Designation Renewal System: DRS by the Numbers. Technical report, Office of Head Start, An Office of the Administration for Children and Families, Washington, DC.

O'Hara, M. W. and Swain, A. M. (1996). Rates and risk of postpartum depression—a meta-analysis. *International Review of Psychiatry*, 8(1):37–54.

Organization for Economic and Co-operation Development Family Database (2019). Parental leave systems. Technical report, OECD Social Policy Division, Directorate of Employment, Labour and Social Affairs.

Pac, J. E., Bartel, A. P., Ruhm, C. J., and Waldfogel, J. (2019). Paid family leave and breastfeeding: Evidence from california. Technical report, National Bureau of Economic Research. NBER Working Paper.

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., and Shah, P. S. (2016). A Systematic Review and Meta-Analysis of a Measure of Staff/Child Interaction Qual-

ity (the Classroom Assessment Scoring System) in Early Childhood Education and Care Settings and Child Outcomes. *PloS One*, 11(12):e0167660.

Persson, P. and Rossin-Slater, M. (2019). When Dad Can Stay Home: Fathers' Workplace Flexibility and Maternal Health. NBER Working Paper w25902, National Bureau of Economic Research, Cambridge, MA.

Pew Research Center (2015). Parenting in America: Outlook, worries, aspirations are strongly linked to financial situation. Technical report.

Pianta, R. and Burchinal, M. (2007). National Center for Research on Early Childhood Education Teacher Professional Development Study (2007-2011). Technical Report 2016-04-12, Inter-University Consortium for Political and Social Research [distributor]. Version Number: v2 type: dataset.

Pianta, R. C., La Paro, K. M., and Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3.* Classroom Assessment Scoring System™: Manual K-3. Paul H Brookes Publishing, Baltimore, MD, US. Pages: xi, 112.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., and Klieme, E. (2014). One Lesson is All You Need? Stability of Instructional Quality Across Lessons. *Learning and Instruction*, 31:2–12.

Rossin-Slater, M., Ruhm, C. J., and Waldfogel, J. (2013). The effects of California's paid family leave program on mothers' leave-taking and subsequent labor market outcomes. *Journal of Policy Analysis and Management*, 32(2):224–245.

Rossin-Slater, M. and Uniat, L. (2019). Paid Family Leave Policies And Population Health. Health Policy Brief, Health Affairs, Bethesda, MD.

Sabol, T. J., Ross, E. C., and Frost, A. (2019). Are All Head Start Classrooms Created Equal? Variation in Classroom Quality Within Head Start Centers and Implications for Accountability Systems. *American Educational Research Journal*, page 0002831219858920.

Saxbe, D., Rossin-Slater, M., and Goldenberg, D. (2018). The transition to parenthood as a critical window for adult health. *American Psychologist*, 73(9):1190–1200.

Thomas, D. V., Fowler, S. A., Cesarone, B., and Rothenberg, D. (2011). The Impact of Publicly Funded Preschool in Illinois: An Analysis of Data from the Illinois Early Childhood Asset Map. Technical Report Technical Report No. 1, The Illinois Early Childhood Asset Map Project, College of Education at University of Illinois at Urbana-Champaign, Champaign, IL.

Thompson, O. (2018). Head Start's Long-Run Impact Evidence From the Program's Introduction. *Journal of Human Resources*, 53(4):1100–1139.

Timpe, B. (2019). The long-run effects of america's first paid maternity leave policy. Working Paper.

Trajkovski, S. et al. (2019). California paid family leave and parental time use. Technical report, Center for Policy Research, Maxwell School, Syracuse University.

von Hippel, P. T. and Hamrock, C. (2019). Do Test Score Gaps Grow before, during, or between the School Years? Measurement Artifacts and What We Can Know in Spite of Them. *Sociological Science*, 6:43–80.

Wang, S., Hu, B. Y., Curby, T., and Fan, X. (2020). Multiple Approaches for Assessing Within-Day Stability in Teacher-Child Interactions. *Early Education and Development*.