Baby bumps in the road: The impact of parenthood on job performance, human capital, and career advancement[*]

Olivia Healy
Cornell University

Jennifer Heissel
Naval Postgraduate School

September 21, 2022

Abstract

This paper explores whether and why a maternal "child penalty" to earnings would emerge even without changes in employment and hours worked. Using a matched event study design, we trace monthly changes in determinants of wages (job performance, human capital accumulation, and promotions). Data come from a usefully unusual setting with required multiyear employment and detailed personnel data: the United States Marine Corps. Mothers' job performance initially declines and gaps in promotion grow through 24 months postbirth. Fathers' physical fitness performance drops somewhat but recovers. These patterns lead mothers to earn relatively lower wages, even absent changes in employment postbirth.

JEL Classification: J24, J16, J18, J45
Keywords: Parenthood, Child Penalty, Gender Wage Gap, Promotion

[*] Olivia Healy is a postdoctoral fellow in the Brooks School of Public Policy at Cornell University in Ithaca, NY (ojh4@cornell.edu). Jennifer Heissel is an associate professor in the Department of Defense Management at the Naval Postgraduate School in Monterey, CA (jaheisse@nps.edu). The authors are grateful to Maj. Tamara Cordero, Capt. Amanda Henegar, Capt. Michael Larson, Maj. Laura Laurito, Capt. Julie Schumacher, Maj. Genevieve Studer, Capt. Ansley White and all Naval Postgraduate School Marines who provided insight into Marine processes and procedures. We also thank Jonathan Guryan, Matthew Notowidigdo, Diane Schanzenbach, Jonathan Meer, Lisa Kahn, David Figlio, Brian Tracy, and seminar participants at the National Bureau of Economic Research Summer Institute (Children), American Economic Association, Society of Labor Economists, Midwestern Economic Association, Eastern Economic Association, Western Economic Association, Association for Public Policy Analysis and Management, Stanford University, and the Military Health System Research Symposium for helpful comments. We are grateful for funding from the Robert Wood Johnson Foundation's Evidence for Action grant program (#77124); the Institute of Education Sciences' Multidisciplinary Program in Education Sciences (Award R305B140042); and Northwestern University's Graduate Research Grant Program. Any errors are those of the authors. The views expressed in this paper do not reflect the views of the U.S. Department of Defense or the U.S. Navy. Note that the data from the Department of Defense are not available for sharing with the public.

After having a first child, mothers' earnings drop precipitously and remain 20–60% lower than fathers' for the decade that follows (Kleven et al., 2019a,b). This pattern, termed the child penalty, is remarkably consistent across countries and contexts (e.g., Aguilar-Gomez et al., 2019; Andresen and Nix, 2019; Angelov et al., 2016; Barth et al., 2017; Bertrand et al., 2010; Fernández-Kranz et al., 2013; Kleven, 2022 and Kleven et al., 2019a). It also drives most of the overall male-female earnings gap in higher-income countries (Cortes and Pan, 2020; Kleven et al., 2020, 2019b). Prior research shows the child penalty comes from three sources: exits from the labor force, reductions in hours worked (conditional on employment), and lower wages among women relative to men after having a child (Fernández-Kranz et al., 2013; Kleven et al., 2019b).[1] This paper focuses on the wage channel behind the child penalty. We explore whether and how determinants of mothers' and fathers' wages respond differently to childbirth. We leverage a usefully unusual setting: the U.S. Marine Corps. The Marine Corps tracks performance data on all employees, as well as human capital accumulation and increases in job rank. Data are updated monthly, which allows us to trace immediate, month-by-month consequences of parenthood on workers' job performance, human capital accumulation, and career advancement—all of which determine wages in our setting and in the broader labor market. Workers in the Marines sign multiyear job contracts, meaning we isolate impacts largely absent confounding factors like labor market exits, reductions in hours worked, and cross-firm job changes (Kleven et al., 2019b; Laffers and Schmidpeter, 2022). In turn, our empirical analysis addresses the question: if men and women were to stay on the job working similar hours after childbirth, would a child penalty in earnings emerge, and why?

Our findings come from workers in the world's largest employer, the U.S. Department of Defense, whose policies and actions have global reach. Gender-based disparities in work outcomes in this setting are of particular concern because they may limit who advances to higher ranks and contributes to decision-making. Beyond implications for the U.S. military, our data provide a rare opportunity to examine job-relevant outcomes in higher-physicality jobs, which comprise a large share of the U.S. workforce

[1] A longstanding literature in economics documents the negative impact of fertility on maternal employment, hours worked, and wages (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). The recent child penalty literature shows that adjustments on these margins among women (and not men) create a mother-father earnings gap that persists over the life course. *Why* only women adjust in these ways after having a baby remains an open question.

and are less studied in the current literature. To contextualize the specialized degree of health and fitness required among those in our sample, Figure 1 presents data on the physical strength and stamina required for work among (1) male and female Marines, (2) those in the top 10 most common female occupations, and (3) those studied in prior literature on the job productivity and promotion consequences of parenthood. Values shown for our sample of Marine men and women reflect the median O*NET level of dynamic strength and stamina of Marine jobs mapped to their civilian equivalents.[2] The most common job for civilian women (registered nurse) requires more strength and stamina than the median Marine woman's job, and many of the other common jobs among women (e.g., elementary and middle school teacher, retail sales supervisor, and cashier) also require some level of physicality. Most prior research on postbirth productivity has focused on the impacts of parenthood among individuals in less-physical jobs, including economics professors (Antecol et al., 2018), lawyers (Azmat and Ferrer, 2017), professionals with a master's degree in business administration (MBA grads; Bertrand et al., 2010), and scientists (Kim and Moser, 2021).

Our outcomes include job performance, human capital accumulation, and career advancement. We measure job performance using scores from job-related physical fitness tests, supervisor ratings, and evaluations of marksmanship skills. Measures of human capital include the number of months workers received job-specific training and their accumulation of years of formal education. Last, career advancement tracks the number of promotions Marines received over time. Though we study these outcomes in a Marine context, the same factors—job performance, human capital, and promotion—also determine civilian wages in most industries.

Our empirical strategy uses the precise month of childbirth as an exogenous shock to parents' work performance outcomes. We assign "placebo births" to nonparent Marines, those who do not have a child during the study window. Changes in nonparents' outcomes help us account for secular time trends that would have occurred absent parenthood. To assign placebo births, we first match nonparents to parents exactly on tenure (in months) in the Marine Corps, job rank, active duty (full-time) or reservist

[2] O*NET dynamic strength levels capture a worker's ability to exert muscle force repeatedly or continuously over time (O*NET, 2022; Department of Labor, 2022). Dynamic strength levels required for an occupation range from "not relevant" for jobs like Postsecondary Economics Teachers to 66 for dancers. Stamina measures the ability of a worker to extend themselves without getting out of breath over long periods.

(part-time) status, and calendar year of the observation. We further restrict matches based on observable characteristics that best predict the likelihood of a first birth. We compare outcomes between parents and nonparents before and after birth (or placebo birth). Compared to a traditional two-way fixed effects model, the placebo birth strategy better corrects for observable differences between parents and nonparents that might affect parents' postbirth outcome trajectories in the absence of having a child. The matching strategy allows us to study cumulative changes in outcomes (e.g., total months spent in training during pregnancy and postbirth), given that we can align parents and nonparents by event time and measure changes in the outcome relative to the event/placebo event. The matching strategy also allows us to explore subgroup differences based on prepregnancy characteristics, given that we can define prepregnancy for the nonparents. We estimate all models separately for women and men to explore whether parenthood affects men along the margins we study. If men are affected, fathers would not be a good comparison group for mothers.

We find persistent effects of parenthood on work outcomes for mothers. Mothers' physical fitness, job performance ratings, and marksmanship scores decline immediately postbirth, as soon as they are observed after any testing exemptions around pregnancy/birth are lifted. Mothers' job performance and marksmanship scores recover by the child's second birthday, but their physical fitness remains 0.2 standard deviations lower than had they not had a child. Mothers also accumulate fewer months of training relative to nonmothers, with gaps beginning during pregnancy and continuing through two years postbirth. Consistent with these patterns, mothers' promotion trajectories slow during pregnancy, with gaps widening, rather than narrowing, during the two years postbirth. These patterns broadly hold across various subgroups, although the long-run promotion effects are driven by junior enlisted (relative to senior enlisted and officers). Junior enlisted have a rigid points-based promotion system, while senior enlisted and officers have somewhat more leeway in the promotion decision process.

We observe minimal impacts of parenthood on fathers' outcomes. The birth of a child leads to small, short-lived declines in physical performance, job performance, and training one month postbirth. We do not find effects on fathers' years of formal education or accumulated promotions.

In supplementary analyses, we show men who have a recorded medical event (e.g., an injury) that limits their ability to be deployed or take a physical fitness test experience changes in their physical fitness, supervisor ratings of job performance, training, and promotions similar to those of mothers. We take this as suggestive evidence that the health shock of childbirth is at least partially behind observed changes in mothers' performance and promotion outcomes. We also document how the changes in physical fitness, supervisor ratings, and marksmanship translate directly into changes in the points used in the junior enlisted promotion process, directly giving rise to gaps between mothers and fathers in career advancement.

Together, our results show that the immediate period after having a child is uniquely challenging for women who remain on the job as U.S. Marines. Our findings suggest that parenthood's mental and physical strain accrues more acutely to mothers, limiting their ability to perform on the job, develop their human capital, and advance professionally after having a child.

Last, we explore whether changes to the length of paid maternity leave during our study window predict changes in the impact of parenthood on mothers' outcomes. We compare mothers who were eligible for 6 weeks of leave, 6 weeks plus 12 flexible weeks of leave after returning to work, or 18 weeks of leave based on the timing of their birth. We do not find statistically significant differences in the magnitude of parenthood impacts across the three policies. However, the estimates are noisy and sometimes indicate worse outcomes for the flexible policy.

Several related studies document adverse productivity shocks associated with parenthood. Similar to our work, Kim and Moser (2021) find female scientists in the 1950s patented less during their childbearing years, relative to fathers and other women, while Azmat and Ferrer (2017) show female lawyers with young children bill fewer annual hours than male lawyers with young children. Gallen (2018) explores how firm-level output in Denmark varies by the gender and parental status of employees in private firms. She finds that mothers are less productive than other workers using a firm-level production function model, particularly during their childbearing years. Our paper focuses on precursors to work output and productivity—job performance and skill accumulation—at the individual level. We build on (Healy and Heissel, 2022) to show that parenthood impacts a range of mothers' performance and human capital outcomes, and that these impacts give rise to gender disparities in career advancement following the transition to parenthood.

A key contribution of our paper is the ability to trace month-by-month responses of men and women in the immediate two years following a first birth. Prior papers on child penalties use yearly measures of earnings or income and cannot detect initial, within-year impacts (Andresen and Nix, 2019; Angelov et al., 2016; Barth et al., 2017; Kleven et al., 2019a,b). We show that parenthood begins to impact women's performance starting *within* the first year after birth across various job-relevant outcomes. Results support the notion that immediate postbirth declines in mothers' ability to perform at work and advance their job skills may lead mothers to exit the labor market, reduce hours worked, or garner lower wages by the time these outcomes are measured one or more years postbirth in other papers. We also show that fathers initially have some decrease in physical performance, but this does not change their promotion rates. Our results demonstrate that even absent changes in employment and hours worked, having a child can create gender-based earnings penalties through its impact on mothers' career advancement.

1 Institutional Background

The DoD employs 2.3 million active-duty and reserve service members. We focus on the U.S. Marine Corps, a service branch within the DoD, where administrative records on job performance are readily available. Marines are an immediate response force, ready to deploy quickly to support combat missions on sea or land. Marines make up 15% of active-duty forces, and the majority are male (92%; Department of Defense, 2018). Given that women are less likely to serve, any findings of declines in work outcomes for female service members due to parenthood may be especially noteworthy.

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.[3] There are over 35 career fields in the Marines, and each has dozens of specializations, referred to as Military Occupational Specialties. Some career

[3] Enlisted service members must have a high school degree and be between 18 and 29 years old when they begin; the Marine Corps is 89% enlisted. Officers must have at least a bachelor's degree. Most Marines are of prime childbearing age, with 81% under 30 years old. More than a quarter are parents (Department of Defense, 2018).

fields are specific to the military (e.g., infantry), and others are also present in the civilian labor market, such as food services, financial management, police and corrections, and legal services.

A service member's occupational specialty and their assigned unit determines their day-to-day work environment. Our analytic sample predominantly consists of active-duty Marines who work full time, Monday through Friday. For these individuals, the workday typically begins with early morning physical training, 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. Our sample also includes reservists, whose service requires participating in training drills one weekend per month and attending a two-week program each year. Reservists typically have civilian careers or enroll in higher education while they fulfill part-time Marine Corps duties. Reservists can be called for active duty, at which point they are active reserves and work as a Marine full-time. We group active reservists with active duty service members for the purposes of this study.

Marines sign a legally binding contract that outlines their required length of service. Initial enlisted contracts typically require four years of active-duty service, while officer commitments are typically three years.[4] These contracts limit Marines' ability to exit the labor force after they have a baby depending on the length of the contract that remains.

Effective job performance in the Marines requires both mental and physical acuity. The Marine Corps uses a standardized set of measures to evaluate performance among both active-duty and reserve Marines. Performance measures include physical fitness tests, supervisor-scored job proficiency, and marksmanship tests. Scores on these measures determine promotions, with different weights placed on each depending on the given promotion juncture. Marines also go into a long-term training duty status when they are selected for certain training courses (e.g., professional military education); training can last several months. Outside of job-specific training, Marines commonly pursue educational degrees while in service, either paid for as part of their job (e.g., being given orders to obtain a master's degree

[4] Active duty contracts can stipulate additional service and require additional years of service in the reserves. At the end of each contract, Marines can decide whether to re-enlist, which involves another contract with a time commitment. About 75% of Marines only complete one contract (U.S. Marine Corps, 2021).

full-time) or paid for privately (e.g., an enlisted Marine pursuing a bachelor's degree outside of work).[5] Training and education also contribute to promotion decisions. Based on these clearly identified measures, Marines can determine what they need to advance and have especially strong incentives to perform well on assessments.[6]

The DoD provides several family-friendly benefits, including subsidized childcare and fully paid parental leave. In a supplementary analysis, we focus on policy changes to the length of paid leave for primary caregivers (most often women).[7]

2 Data

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

Our first set of outcomes consists of the three primary measures of job performance used for promotion and retention decisions. First, physical performance captures Marines' ability to perform physically demanding work tasks. Marines take two standardized tests per year: the Physical Fitness Test (PFT; timed running, crunches, and pull-ups/push-ups) in the first half of the year (typically May–June) and the Combat Fitness Test (CFT; timed running, a combat-related obstacle course, and upper-body strength measured by ammunition can lifts) in the second half of the year (typically October–December). Generally, an individual's command gives a Marine a few weeks' notice that they will run a PFT/CFT on a given day, but competing priorities mean some portion of Marines have to do it at a later time.[8] Scores

[5] Marines who obtain education as part of their job add time to their commitment as part of a "pay-back tour."

[6] For enlisted (officers), E1 (O1) is the lowest rank and E9 (O10) is the highest. A composite score based on performance metrics determines promotion through E5, conditional on meeting requirements for minimum time in service and time in the current rank. Promotions at lower ranks are relatively automatic after a given number of months in service and months in rank, while promotion to E4 and E5 are more competitive (Larger, 2017; U.S. Marine Corps, 2017). For promotion at ranks above E5, the same performance metrics are reviewed by an evaluation board to determine promotion. Both Marine Corps and civilian promotions are based on work performance, but the Marine Corps promotion system is perhaps more systematic.

[7] Paternity leave changed from 10 to 14 days during this time. We do not examine paternity leave in this paper.

[8] Appendix Figure A.1 displays the distribution of physical fitness test timing by month and parenthood status. There is

are awarded on a 300-point scale, which is adjusted for age and gender such that women do not need to do as many pull-ups as men, and older service members do not need to run as fast as younger ones to achieve the same score. We standardize points-based physical fitness scores by year, gender, and test type. We combine the Z-scores for the two tests into one measure, generally observed twice per year per Marine. During our study window, women were not required to take the test when pregnant and 6 months postbirth. We resume measurement of physical performance for women at 8 months postbirth due to concerns that commanding officers may allow some women who are at 7 months postbirth in December to skip the CFT that cycle.

The second job performance outcome we measure consists of supervisor evaluations of Marines' work performance. Supervisors regularly assess Marines' using one of two rating scales, depending on the Marine's rank. During our study window, junior enlisted received proficiency and conduct marks ("ProCons"), and senior enlisted and officers receive Fitness Reports ("FitReps"). Both assessments require supervisors to assess a Marine's performance across a range of professional domains, such as technical knowledge, effectiveness, and communication skills.Marines receive these evaluations even if they are on leave; performance on a given assessment is then based on the time from their last evaluation until they went on leave. We standardize supervisor ratings by year, gender, and assessment type. We combine the Z-scores into one outcome we call job performance. We generally observe ratings at least twice per year among junior enlisted, once in the first half and once in the second half. For senior enlisted and officers, we observe supervisor ratings at least once per year. If a Marine is transferred, discharged, or promoted, or if their supervisor changes, they receive additional performance ratings.[9] We are missing supervisor ratings for junior enlisted who left the Marines before October 2017. We observe the full history of performance ratings (including prior to October 2017) for any Marine who was active as of October 2017. The subjective nature of supervisor ratings means we cannot distinguish true changes in

more variation across months than across groups. Mothers are slightly more likely to take a test in June than other groups, either because they have waivers before that or because they otherwise delay the test. They are required to take the PFT by June and CFT by December; if all mothers took the PFT in June we would still have a distribution of scores across months relative to birth, which is what we focus on. An F-test of the mean month of the test does not indicate mothers statistically take the test later than other groups ($F(3, 26, 557) = 1.23$, p-value $= 0.296$, with standard errors adjusted for clusters from the main analysis).

[9] 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).

job performance from supervisors' *perceptions* of changes in performance using this measure.

Our third job performance outcome captures rifle and pistol marksmanship assessment scores. Marksmanship assessments evaluate Marines on their target shooting skills and award points according to performance. While this measure is clearly a firm-specific measure of job proficiency, strong marksmanship performance is partially cognitive and requires practice, focus, and concentration. We standardize marksmanship scores by year, gender, and weapon (rifle/pistol). The marksmanship assessment is required once a year at junior levels and becomes optional for more senior Marines. Those who perform at the highest level (rated "expert") are exempt from re-testing the following year, making outcome data on this measure sparse. Like physical performance tests, the marksmanship test requirement is waived for pregnant mothers and not assessed when mothers are on parental leave. We resume measurement of marksmanship scores for mothers starting 5 months postbirth, given that some mothers in our study window could have been on leave for 18 weeks.

Figure 2 displays counts of physical performance, job performance ratings, and marksmanship scores relative to birth for mothers; the figure shows a drop in the frequency of observations around the birth event for physical performance and marksmanship scores, corresponding to periods during which women are exempt from tests. We exclude unshaded observations of women's physical performance and marksmanship scores from our analyses.

Our second set of outcomes captures human capital development. To measure firm-specific human capital, we observe whether a Marine is in training each month. We create a cumulative count of months of training observed for each Marine relative to 10 months before they have a child. A value of 1 on the variable indicates 1 additional month of training achieved since just before pregnancy. We under-count total training, because we only capture one type of training status. In general, a Marine's command decides whether, where, and when an individual will go to training based on the needs of the Marine Corps. Training could be located out of state, and dependents would stay behind at the original base location while the Marine attends the training. The mean training spell lasts 5 months (IQR = 2–6 months).

To measure general human capital, we track increases in Marines' formal education levels, which increase as they gain credits from institutions of higher education. This is reported as years of education (e.g., an associate degree is 14 years of education). Increases in formal education are considered in the promotion process, so Marines are incentivized to keep this information updated. While the Marine Corps may send officers to obtain master's degrees to fit its needs, in general educational attainment among enlisted are initiated by the individual on their own time.

Finally, we track promotions over time. We observe each Marine's job rank on a monthly basis and can trace successive increases in rank over time. We count the number of increases in job rank (i.e., promotions) a Marine achieves relative to 10 months before they have a child. A value of 1 on the variable indicates 1 promotion received since just before pregnancy.

Our preferred sample uses a semi-balanced panel of observations to ensure our results are not driven by selective attrition. We require first-time parents (and their potential matches) be observed continuously in the sample for 12 months prior to birth and 24 months after (n=2,801 mothers). We also require parents to have a military entrance exam score (AFQT or GCT), at least one observed pre-birth fitness score and for parents to have at least one potential nonparent match with the same months of service, rank, and reserve status in the same prepregnancy calendar year. Among mothers, this reduces our sample to n=2,492.

We rely on the Standard Occupational Classification (SOC) system from O*NET, 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, while first-time Marine mothers work in sales or office roles. A small share of first-time Marine parents in our sample are officers (akin to civilian managers): 7% and 14% of Marine mothers and fathers, respectively. Results from our analyses may generalize best to younger workers with low levels of formal education.[10]

[10] Appendix Table A.1 provides details on Marine parents and employed civilian parents with a first child under age 1.

3 Empirical Approach

The ideal experiment to isolate the causal effect of fertility on men's and women's work performance would randomly assign pregnancy and parenthood to workers. Random assignment would ensure that (on average) differences in outcomes were not driven by underlying characteristics of the types of people who chose to become parents but rather by the transition to parenthood itself. Of course, random assignment of childbirth 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 differ from nonparents in ways that might also correlate with work performance.

We use the timing of first birth to identify the effect of the transition to parenthood. If the transition to parenthood affects work outcomes, then birth should generate a sharp change in these outcomes at predictable time points. We can attribute any discontinuity in the outcomes at those time points to the pregnancy or birth itself if we assume that other factors that shape job outcomes do not also undergo a sharp change at those same times. In other words, while the choice to have a child may be endogenous, the exact timing of conception and subsequent childbirth serves as a shock to the outcomes of interest.

We employ a version of a two-way fixed effects (TWFE) event study strategy using the shock and precise timing of birth event. Our goal is to minimize two sources of bias. The first source of bias, which we call TWFE bias, arises when the timing of treatment varies (as it does in our setting) and already-treated units serve as comparison cases for later-treated units, contaminating estimates of counterfactual time trends.[11] To address TWFE bias, Callaway and Sant'Anna (2021) and Sun and Abraham (2021) propose isolating cohorts of units all treated at the same point in time and selecting comparison cases for each treatment time group that *exclude* already-treated units. This strategy functionally aligns event-time and calendar time across treatment and control cases within each treatment time group.

[11] A traditional TWFE approach incorporates time and unit fixed effects to estimate post-treatment impacts (whether dynamic or constant). Already-treated, not-yet-treated, and never-treated units (if included) contribute to estimates of counterfactual time trends, or the time fixed effects. The estimate of the causal parameter under a traditional TWFE model is then a weighted average of all of two-group/two-period difference-in-difference estimators in the data. As treatment timing gets later, more of the two-group/two-period difference-in-difference estimators compare the change in a unit moving from untreated to treated against units who are treated in both the before and after period. This statistical fact requires stronger assumptions than previously recognized, with both a parallel trends assumption *and* an assumption of consistent treatment effects over time and units needed for identification (Goodman-Bacon, 2021). Baker et al. (2021) show that not accounting for the bias in settings that do not meet these assumptions can affect the point estimates of the policy or treatment in question.

The second source of bias, which we call counterfactual bias, arises if the chosen comparison units do not effectively approximate counterfactual trends in outcomes. In our setting, not all Marines without a birth (or yet to have a birth) would provide a good estimate of counterfactual time trends for those who do have a first birth, especially for promotion outcomes. Promotion eligibility and timing depend on a Marine's current job level and time in service. A second-year enlisted Marine becomes eligible for promotion sooner and can promote faster than a fifth-year officer. Active duty vs. reserve contexts also give rise to different promotion probabilities. As a result, we face large counterfactual bias if we do not thoughtfully select comparison units for estimates of promotion effects. The same concern around counterfactual bias holds true for our estimates of other work outcomes, given that professional expectations for job performance and human capital development depend (less formally) on job rank, tenure, and active duty/reserve contexts.

Ideally, our empirical strategy would isolate groups of parents with first births in the same month-year (to address TWFE bias) and then draw never-treated comparison nonparents with the same job tenure, job rank, and reserve status as parents (to address counterfactual bias). Under this approach, some groups of parents have few to no nonparent comparison units (especially among reservists and higher ranks). Nonparent comparison units also vary from parents on other important dimensions (e.g., average physical performance, education, age) not defined in the exact match requirements, which generates new concerns around counterfactual bias.

Given this tension, we require that parents and nonparents match on tenure, rank, and reserve status, but we do not require exact treatment time matching. Instead, we connect parents to nonparents within the same calendar year. Parents' and nonparents' birth/placebo birth events do not occur in precisely the same month-year, but all observed outcomes occur within the same 12-month timeframe. We hypothesize, in our context, that this method provides the best counterfactual to the treated group. In other words, we are willing to risk some TWFE bias to reduce counterfactual bias. This tradeoff is common in settings where there are a limited number of never-treated or not-yet-treated comparison units once the researcher isolates each cohort of units treated at the same time.[12]

[12] Appendix B explores alternative specifications: a standard TWFE model, a stacked TWFE model that exact matches on calendar month, and a placebo event study without binned end points. The broad takeaways are consistent across all models,

3.1 Assigning Nonparents to Placebo Births

We use adaptive ridge LASSO (least absolute shrinkage and selection operator) with 10-fold validation to select the best predictors of a first birth in our sample. Possible predictors include months of service, an indicator for officer (relative to non-officers), an indicator for reservists (relative to active duty), year, age, race/ethnicity, AFQT scores[13], marital status, an indicator for whether a spouse is also in the military, years of education, occupational groups, most recent physical performance score, and interactions among all variables. We restrict the sample to parents and potential matches we would be able to observe at least 12 months before and 24 months after the match. We run this predictive model separately among women and men, measuring all characteristics of first-time parents 10 months before an observed birth (prepregnancy). We then obtain a predicted propensity score for each parent and nonparent.

Among groups of parents and nonparents with exact matches on number of months of service, job rank, reserve status, and observation calendar year, we select up to five nonparents for every parent that are closest in terms of their predicted propensity to have a baby 10 months later.[14] We assign nonparents to a placebo birth event 10 months after the time of the match. Analyses then compare the changes in outcomes for first-time parents to the average change in outcomes for up to 5 most observably similar nonparents to whom they match. Each parent receives a weight of 1 in the analysis, while each nonparent receives a weight of 0.2 per match-month in the case where 5 distinct nonparents match to each parent.

Table 1 displays descriptive characteristics of parents and matched nonparents separately by gender. We require matches to be exact on months of service, job rank, reserve status, and calendar year, so the groups exactly match on this first set of variables. The next set of variables are not exact matches but are

but the size of the point estimates differ by model and highlight the importance of choosing the best comparison group. Promotion is mechanically tied to rank and time in service. We would not want to match a low-ranking officer to mid-ranking enlisted, even if they have similar rates of promotion in the preperiod, as subsequent expected promotion trajectories differ even in the absence of a child. Indeed, the promotion gap is 55% larger for mothers in the stacked TWFE model than the preferred model; for fathers, the sign flips direction and loses statistical significance. Our approach prioritizes defining an appropriate counterfactual while still considering how to minimize TWFE bias. This strategy may be useful in settings where other variables (e.g., months of service) are particularly important to consider for creating a counterfactual group.

[13] We have missing AFQT score data for many officers, who generally take GCT exams. We use a cubic model of standardized GCT scores to predict standardized AFQTs for those with both scores, then use this model to predict AFQT scores for those with only GCT scores.

[14] We conduct nearest-neighbor matching with replacement, meaning the same nonparent can be matched to different first-time parents, or parents can match to less than 5 nonparents, who then get higher weights. About 2.8% of mothers and 0.5% of fathers match to fewer than 5 nonparents.

included in the LASSO model predicting likelihood of first birth. First-time parents and nonparents with placebo births look almost identical, with a few small differences within gender. These differences are functionally small (e.g., mothers are 22.57 years old while their placebos are 22.70 years old).

3.2 Flexible Event Study Estimation

We create a series of event study datasets across exact-match groups (defined by month of service, job rank, reserve status, and calendar year of the pregnancy/placebo pregnancy). We include all month-years of data from before through after the birth/placebo birth, then stack these exact-match-group datasets.[15] Nonparents assigned placebo births approximate counterfactual time trends in outcomes that first-time parents would have experienced, assuming outcomes would have evolved similarly between parents and those with placebo births. An unbiased estimate in this setting requires an assumption of conditional parallel trends between parents and placebos, which is a weaker assumption than unconditional parallel trends had we drawn on all nonparents from the full sample (Roth et al., 2022).

We estimate a fully flexible event study specification separately for men and women as follows:

$$Y_{igtr} = \sum_{r=k_{min}}^{k_{max}} \mathbb{1}(t = t^*_{ig} + r)\theta_r + \pi P_i + \sum_{r=k_{min}}^{k_{max}} \mathbb{1}\left[(t = t^*_{ig} + r) \times P_i\right] \beta_r + \alpha_g + \phi_t + \varepsilon_{igtr} \qquad (1)$$

where t^*_{ig} is the month-year of the real or placebo birth for individual i in match group g based on calendar time t. We measure month relative to birth as r. Coefficients θ_r estimate changes in the outcome for each month r after birth (or before, if $r < 0$). This is analogous to event time fixed effects, estimated among both parents and nonparent matches. P_i is a binary variable equal to one for all first-time parents, and we expect π to be zero given that parents and placebos are similar in the preperiod. Then, β_r represents how much the parents differ from their placebos at a particular time relative to birth. Effects are measured relative to month $r = -10$, which corresponds to 10 months prior to birth and approximately 1 month before the start of the pregnancy for the parents. We focus on month-by-month effects starting 24 months

[15] Each month-year observed for a given parent will appear exactly once as they only have one first birth. A given month-year for nonparents may occur multiple times if the same non-parent is matched in different month of service/rank/reserve status/year cells. This matched individual would have different relative event-time points, defined by the time point of their match and assigned placebo birth, for the same observations of calendar month-year.

before birth through 24 months after for most outcomes; for job performance ratings we instead focus on effects starting at $r = -18$ because the evaluation requires workers to be on the job for roughly 6 months before it is typically assessed. We bin event time endpoints below $r = -24$ (or $r = -18$ for job performance) and above $r = 24$. Including binned event time endpoints allows us to estimate time fixed effects ϕ_t to account for month-by-year changes over time in the outcome (e.g., changes in fitness test standards in a particular year) separately from event time fixed effects. Excluding $|r| > 24$ and dropping the binned end points produces nearly identical results (see Appendix B). We include α_g to create a within-match-group comparison and ε_{igt} as the error term. Thus, θ_{12} represents the average outcome 12 months after birth for nonparents, relative to $r = -10$, while β_{12} estimates whether this change is more positive, more negative, or the same for parents.

Eq. 1 allows us to estimate prepregnancy differences in outcomes between first-time parents and nonparents, reflected by β_r estimates when $r < -10$. Our placebo birth assignment procedure does not require outcomes between parents and nonparents to move together when $r < -10$, but we expect β_r for $r = [-24, -11]$ to be zero if nonparents provide a suitable comparison to parents. We show evidence that parents and nonparents are on parallel trends, lending confidence to this estimation strategy.

3.3 Incorporating Linear Splines

We can improve precision in our estimates of the evolution of outcomes by using a semi-parametric spline specification, especially for estimates with smaller subsamples and outcomes that are not observed every month. Our goal is to identify any level shifts in outcomes during pregnancy, trends during pregnancy, level shifts immediately following birth, and any recovery trends following the immediate impact of birth. Similar to Lafortune et al. (2018) and Bailey et al. (2021), we create a more parsimonious model of changes in outcomes over time by defining:

$Pregnancy_{igtr} = 1$ during months relative to birth $r = [-9, -1]$ and 0 otherwise (for an intercept shift during pregnancy);

$PregnancyTrend_{igtr} = [0, 8]$ corresponding to $r = [-9, -1]$ and 0 otherwise (for monthly trends beyond the intercept shift during pregnancy);

$Postbirth_{igtr} = 1$ during months relative to birth $r = [q, 25]$, where $q = 1$ or the earliest time point when the outcome is able to be assessed again after birth (e.g., starting 8 months after birth for mothers' fitness tests)[16], and 0 otherwise (for an intercept shift following birth);

$Recovery_{igtr} = [q + 1, 25]$ corresponding to $r = [q + 1, 25]$ and 0 otherwise (for monthly trends beyond the intercept shift after birth); and

$\Delta Recovery_{igtr} = [1, 13]$ corresponding to $r = [13, 25]$ and 0 otherwise (for any change to the monthly recovery rate that begins at 13 months).

We modify Eq. 1 to estimate a semi-dynamic specification using these spline parameters:

$$\begin{aligned} Y_{igtr} =& \theta_0 Pregnancy_{igtr} + \theta_1 PregnancyTrend_{igtr} + \theta_2 Postbirth_{igtr} + \theta_3 Recovery_{igtr} \\ &+ \theta_4 \Delta Recovery_{igtr} + \pi P_i + \beta_0 (Pregnancy_{igtr} \times P_i) + \beta_1 (PregnancyTrend_{igtr} \times P_i) \\ &+ \beta_2 (Postbirth_{igtr} \times P_i) + \beta_3 (Recovery_{igtr} \times P_i) + \beta_4 (\Delta Recovery_{igtr} \times P_i) \\ &+ \alpha_g + \phi_t + X_{igtr}\gamma_j + (X_{igtr} \times P_i)\delta_j + \varepsilon_{igtr} \end{aligned} \tag{2}$$

where effects are measured relative to the prepregnancy average ($r \leqslant -10$), similar to Borusyak et al. (2021).[17] Parents and their matches contribute to all coefficient estimates θ_j, while slope parameters β_j are specific to parents and captures any change above and beyond that of nonparents. We present a diagram of this model in Figure 3. We use this semi-dynamic spline specification to estimate postbirth effects for men and women at key time points (e.g., 12 and 24 months postbirth). The vector X_{igtr} includes two binary indicators for binned event time endpoints, one for event times $r < -24$ (or $r < -18$ for job performance) and event times $r > 24$ to mirror our estimation strategy from Eq. 1.

Abadie and Spiess (2021) show that clustering should be at the match-group level when doing matching without replacement to account for within-group correlation induced by the matching procedure.

[16] We define q starting at $q = 1$, the month after birth, due to ambiguity about whether outcomes measured in the month of birth itself $r = 0$ reflect pre- or postbirth measures.

[17] Using all pretreatment periods is more efficient than using only the period just before the event takes place, but it is also more biased if parallel trends do not hold (de Chaisemartin and D'Haultfœuille, 2021). The event study estimates from Eq. 1 use the period just before treatment as the reference group and thus avoid this issue.

That does not solve the problem in matching with replacement that individuals' outcomes are correlated if they are part of multiple matched groups. To address this latter concern, we include two-way clustering at the individual and match-group level.

4 Results

4.1 Job Performance

Figure 4 presents results from our flexible event study model estimated using Eq. 1. We first examine whether outcomes evolve smoothly leading into pregnancy, suggesting that placebo parents provide a suitable counterfactual estimate of general time trends for the event study sample. The bottom left of each panel includes the p-value of an F-test of whether the prepregnancy point estimates jointly equal zero. We begin with job-related physical performance scores (Panel A). The prebirth period is not jointly statistically different from 0 (p-value of an F-test of joint significance=0.151), indicating that before giving birth mothers' trajectories did not differ from the nonparent women. We exclude outcomes for women during pregnancy and seven months after birth, as policies allowed women to opt out of the physical assessments while pregnant and postbirth. Once women take the test postbirth, their performance declines are large and persistent. Even 24 months postbirth, women's physical performance scores are lower than expected, after accounting for general time trends using the nonparent women. For men, performance declines begin during the mother's pregnancy and reach their lowest point 1 month postbirth. The declines are short-lived.

Panel B shows some evidence of lower supervisor ratings of job performance for women in the 2 years postbirth, though estimates are noisier than physical performance given that we are missing observations for those who left the Marines before 2017. Having a child does not appear to affect fathers' supervisor-rated job performance.

Finally, Panel C indicates that parenthood is unrelated to marksmanship for mothers. Women are exempt from these assessments during pregnancy and while on leave following birth. Given the rarity of this assessment, the estimates are noisy. If anything, marksmanship improves postbirth for fathers.

Table 2 displays the spline results using Eq. 2. This model has the advantage of smoothing the estimates using data in nearby time points, which is particularly helpful for the noisier event study estimates where outcomes are not observed monthly (i.e., physical performance, job performance and marksmanship). Rather than present coefficients for each spline parameter (e.g., the slope of the postbirth period), we instead predict effects at various time points using coefficients from the model.[18] This can be interpreted as how much parents changed from the prebirth period to the given point in time, net of expected secular trends in the outcome, and whether this change statistically differs from zero.

Beginning at 8 months postbirth, mothers' physical performance is 0.50 standard deviations below their expected average. Mothers recover somewhat, and by 12 months postbirth their predicted physical performance scores are 0.29 standard deviations below expectations. Mothers' physical fitness recovery slows in the child's second year of life. Two years after having a baby, mothers' predicted physical performance remains 0.18 standard deviations lower than before the pregnancy.[19]

Supervisor ratings do not change during pregnancy (see Appendix Table A.2). One month postbirth, mothers score 0.17 standard deviations lower than expected, relative to changes in the placebos. This difference is 0.18 standard deviations at one year postbirth and a non-statistically significant 0.07 standard deviations by 24 months postbirth. Marksmanship is lower for mothers than their placebos postbirth, though by 24 months postbirth the difference is no longer statistically significant.

The patterns for fathers' job-relevant physical performance is consistent with mothers' but smaller in magnitude (Panel B of Table 2). Fathers' scores begin falling during pregnancy (see Appendix Table A.2), then drop to 0.12 standard deviations below expectations in the month after birth. By 12 months postbirth, the fathers are 0.05 standard deviations below expectations, and the effect is a precise zero by 24 months postbirth. For job performance, scores are slightly negative in the first month postbirth, but return to match the placebos by the child's first birthday. For marksmanship, if anything fathers may improve over time.

[18] We include the slope parameters in Appendix Table A.2 for reference.

[19] Physical performance assesses a combination of cardiovascular health, endurance, and strength. Appendix Table A.3 presents raw scores for each item on the fitness assessments. Mothers run more slowly (i.e., run times increase) and complete fewer crunches, pull-ups, and lifts.

4.2 Training, Education, and Promotion

We next turn to human capital development and promotion outcomes. These outcomes are recorded monthly, meaning every sample member has an observation for every month-year, giving us more precision in our estimates. Figure 5 presents results from Eq. 1; Figure A.2 displays unadjusted weighted means over time for parents and their placebos.

Figure 5 Panel A shows effects on months of job-specific training, where the Y-axis measures the total count of months in training relative to $r = -10$. Prepregnancy point estimates do not jointly differ from zero, but there appears to be a downward trend, particularly among mothers. Yet, starting in pregnancy mothers' accumulation of months spent in training slows even more relative to nonmothers'. The gap does not close after birth. Table 2 shows that the gap grows from 0.46 months immediately after birth to 0.84 months 24 months postbirth, which is about a 51.9% difference from the nonparent mean growth of 1.61 months in $r = [-10, 24]$. There does not appear to be a meaningful impact of birth on accumulated job-specific training among fathers.[20]

Panel B shows the impact of a first birth on total years of formal education, a transferable measure of human capital. Mothers have slightly lower educational attainment than their placebos by 24 months postbirth, but there is some evidence of pretrends in Figure 5; when we control for pretrends in Appendix Table A.4 we do not observe any impact of a birth on mothers' total years of education. Among fathers, there is a statistically significant but practically small increase relative to the placebos of 0.01 years of education 24 months following birth. However, we do not observe this effect when we control for pretrends in the outcome. The magnitude of the education effect is small in all models. Our overall conclusion is that birth has minimal effects on working parents' educational attainment in our context.

The final row in Figure 5 displays promotions, which is a cumulative count relative to $r = -10$. Mothers and their placebos move together before pregnancy, but a gap emerges around the time of birth. Mothers never catch up to their placebos; if anything the gap grows over time. From Table 2, the

[20] To address concerns around pretrends, we control for a linear pretrend and show predicted impacts in Appendix Table A.4. The magnitude of the effects for months of training among mothers is smaller. Among fathers, controlling for pretrends produces results that suggest fathers' time spent in training increases relative to nonfathers after birth. We interpret training results with caution, given that estimates are sensitive to the specification.

promotion gap is 0.03 at $r = 1$ but grows to 0.09 promotions by $r = 24$. The placebos averaged 1.31 promotions in $r = [-10, 24]$, so this is an 6.7% reduction. Fathers are statistically about the same as their placebos on this outcome.

We conduct a placebo analysis to confirm the results are not mechanically created by the matching algorithm in Appendix Table A.5. All results are statistically indistinguishable from zero, as expected.[21]

4.3 Heterogeneity across Subgroups

Subgroup analyses inform whether the impacts of parenthood on job outcomes are broad-based or concentrated among specific groups in ways that might point at the underlying mechanisms driving our results. We focus first on differences in effects between reserve and active-duty service members. Most reservists have full-time civilian jobs but the same physical and job performance requirements as active-duty. If we see larger (or smaller) impacts of birth among reservists, it may give a sense for how our findings would translate to civilians as compared to those who make their primary living being in the military.[22] We then consider variation in the impacts of parenthood based on whether the Marine parent works in an especially physically demanding job. All Marines are expected to maintain job-related physical fitness, but those in more physically demanding roles may have different incentives to stay fit or more ground to lose in terms of their physical ability after childbirth.[23] Next, we explore impacts among unmarried parents, those with potentially more caregiving responsibilities and less support at home. Our last three subgroups of interest include: (1) junior enlisted who may be less attached to their jobs than senior enlisted or officers, less able to advocate for themselves, and more directly affected by the rigid

[21] To conduct the placebo analysis, we remove all mothers from the sample and randomly assign women from the potential matching pool as placebo parents. We run the LASSO model and matching process to identify matches to these placebos, then run the main analysis. The pretend mothers do not differ from their matches.

[22] We do not have the power to conduct the reservist subgroup analyses for physical performance, job performance, and marksmanship scores for women due to only a small number of female reservists and sporadic nature of the assessments.

[23] We categorize Marines as working in a high-physicality ($> physical$) or low-physicality ($< physical$) job based on whether their job responsibilities place them in the top or bottom half of our sample's distribution of O*NET's dynamic strength index (O*NET, 2022). We exclude individuals working in jobs with no link to an O*NET classification (about 9% of mothers and 32% of fathers). The median mother on the physicality scale in the low-physicality group is equivalent to an office clerk or cartographer; the median father in the low-physicality group is equivalent to a musician or administrative services manager. For both mothers and fathers, the median in the high-physicality group is equivalent to an aircraft mechanic or postal clerk.

points-based promotion system; (2) parents who have a second child soon after the birth of their first child who may drive longer-run effects 24 months postbirth; and (3) Marines who leave the military within three years after having a child and who may not be less invested in recovering their performance.[24]

We conduct subgroup analyses by interacting an indicator variable for a characteristic (e.g., reservist) with the variables from Eq. 2 to measure whether the magnitude of the parenthood effect differs by subgroup type.[25] In general, we lack power to test small differences between subgroups of mothers but are well-powered to detect even small differences among subgroups of fathers. We present results from heterogeneity analyses in Figure 6, displaying physical fitness, training, education, and promotion outcomes. We are underpowered to detect subgroup differences in job performance ratings and marksmanship scores and present these results in Appendix Figure A.3. Cross-group differences that differ at the 1% level have filled-in markers in Figures 6 and A.3. We focus our discussion on subgroup patterns in physical fitness and promotion outcomes, our two most robust results.

Declines in physical performance and promotion among mothers appear largely broad based, with two key exceptions.[26] There appears to be no lasting impact of parenthood (at $r = 24$) on physical fitness and promotion outcomes among (1) senior enlisted/officers and (2) mothers who remain employed 36+ months postbirth. In other words, negative impacts of parenthood on physical ability and career advancement at 24 months postbirth are concentrated among junior enlisted and mothers who leave the military within three years of having a child. It is difficult to interpret the pattern of results for mothers who leave the military within three years of having a child vs. those who stay; mothers who leave may not have strong incentives to perform well on military-specific tasks because they plan to leave, or they may leave *because* of birth-related performance declines. We cannot distinguish between these two

[24] The comparison of parents who have a second birth splits the sample by those who have a second child within 24 months postbirth and those who do not.

[25] Definition of the subgroup is based on the matched parents' characteristics for the placebos, rather than characteristics of placebos themselves. This ensures placebos are in the same group as their matched parents, even if they do not exactly match on a characteristic. We use $r = -10$ to define most subgroups, but use $r = 0$ for marital status because we are most interested in family characteristics when the baby arrives. This time gives couples surprised by a pregnancy some time to marry.

[26] Across subgroups, almost all mothers have lower physical performance when measured again after the birth and 24 months later, but there is some variation in the size of the effects. For example, single mothers, those from more physical jobs, and those who have a second child experience larger declines in their outcomes, but even married mothers, those from less physical jobs, and those who do not have a second child still experience the negative impacts of parenthood at $r = 24$.

possibilities.[27]

Among fathers, we observe lower physical performance immediately postbirth on average, but the size of the impact varies across groups. Reservists, those in more physical jobs, single dads, junior enlisted, and those who stay on the job for less than 36 months after a birth have larger negative physical fitness effects than their counterparts. Moreover, null main effects on training, education, and promotion mask some subgroup differences; for instance, fathers working in less physical jobs experience declines in months of training, increases in years of education, and more promotions at 24 months postbirth. Fathers who remain employed as Marines for 36 months or longer after a birth similarly show increases in months of training, years of education, and number of promotions. There is some indication, then, that among fathers in less physically demanding jobs and among those who remain employed for 3 years after a birth, the impact of parenthood on some job outcomes could be positive.

4.4 Channels Behind Delays in Mothers' Promotion Trajectories

In our setting, workers are promoted as they accumulate tenure and perform their job roles well. At the same time, discrimination can also shape promotion outcomes, given that assessments are sometimes subjective and the decision to promote someone may involve a degree of individual discretion.[28] We take two approaches to begin to disentangle whether changes in performance/skills on the job drive the observed promotion slowdown for mothers in our context, or whether discrimination is at play.

First, we highlight that the negative effect of parenthood on mothers' promotion outcomes is concentrated among more junior workers (junior enlisted) where the promotion system is formulaic (see Figure 6). For junior enlisted, each worker receives promotion points based on how well they score on their job performance assessments, whether they increase their formal education (but not training), and each additional month of tenure they accrue. If a junior enlisted service member's promotion points exceed the designated threshold for their job category in a given quarter of the year, they are promoted. The

[27] In our sample, 25% of parents leave the Marines during the third year of their child's life (25 to 35 months postbirth).

[28] Structural features of the promotion process, rather than individual-level discrimination, may also constrain the odds of promotion for certain groups. For example, assessments may systematically disadvantage women even if they are gender-neutral on their face (e.g., testing abdominal strength of all Marines despite that this will be more difficult for women soon after childbirth).

promotion points formula is known, and our findings show decreases in three of the key inputs in the points calculation (fitness test scores, supervisor ratings, and rifle/pistol marksmanship evaluations) in response to motherhood. Other metrics included in the points calculation are infrequent accomplishments that make individuals eligible for bonus points, such as serving as a recruiter or drill instructor, or ones that cannot or do not decline postbirth (months of tenure and formal education). We can infer, then, that documented changes in measured job performance drive the promotion slowdown we observe among mothers, given that promotion slowdowns occur among the junior enlisted group subject to the promotion points system. In contrast, for senior Marines, the promotion system is less rigid, and there is no average effect of motherhood on promotion. For senior Marines, similar factors play into the promotion decision but the decision comes from discussions among "boards" (akin to panels) of service members rather than a points calculation. We observe drops in senior women's physical fitness and job performance scores similar to that of the junior enlisted (see Figures 6 & A.3), but not a similar slowdown in career advancement. It is possible then, that with more leeway in the decision-making process, senior mothers' career advancement suffers less in response to parenthood.

That we do not observe slowdowns in career advancement among the group of women for whom promotion decisions are more subjective (senior Marine mothers) suggests that discrimination against mothers does not drive our observed change in promotion trajectories for mothers. Yet, some degree of discrimination may still exist. We document declines in supervisor-rated job performance among mothers after childbirth, and these ratings allow for more subjectivity than, say, a timed 3-mile run. Supervisor job performance ratings are also the most heavily weighted input into the promotion points system, and so bias on this measure would have large consequences for promotions since lower job performance ratings would reduce promotion points more than similarly sized declines in physical fitness score, for example.[29] Our second exercise explores this idea more formally. To do so, we estimate the consequences of medical limitations that *men* experience on the job. Observing medically-limited men receiving lower job performance and promotion rates would suggest that it is a medical event, not birth-

[29] The effect for supervisor ratings is smaller than the physical fitness effect in standard deviation units, but supervisor ratings have about twice the impact on promotion points at $r = 8$ than physical tests. See Appendix C for more details on a promotion points simulation.

based discrimination per se, that is associated with lower outcomes.

We capture whether men have either an official "medical" status in their files or miss a fitness test with a medical waiver. These two designations are the same ones used to identify pregnant and postpartum women who do not need to deploy or take a fitness test. Medical conditions outside of pregnancy and birth are infrequent; only 11.5% of Marines who do not have a baby in the study window are ever in a medical status (14.4% of Marines have a baby in the study window). We identify the first appearance of a medical status on file as $r = 0$, and then proceed as in the main parent analysis with identifying a matched group of nonmedical comparisons at $r = -10$. We leave the same 10 months between the match and start of the medical status as we do with pregnancy and childbirth in case there is a delay in recording the medical status measure (e.g., a person is injured, but their command only updates their status when it causes them to miss a fitness test). We include only nonparents in this analysis to avoid contamination with parenthood effects. We focus on men to obtain a large enough sample of medical events and avoid the potential for gender bias against women.

Figure 7 displays the results. We find men whose records show a documented medical status perform 0.2 standard deviations worse on physical fitness tests initially following the medical event; this is an underestimate of the effect given that those with more extreme medical limitations are exempt from physical fitness tests initially after the medical event. The impact of going on medical status dissipates in magnitude but remains around 0.1 standard deviations lower at $r = 24$. Job performance, months of training, and promotions also decline for men who have a medical event. Twenty-four months after we observe the medical status turn on, men are rated 0.1 SD lower on job performance evaluations, accumulate 0.6 fewer months of training, and achieve 0.07 fewer promotions relative to their matched comparisons. The magnitude of these effects is very similar to the estimated effects at $r = 24$ postbirth for mothers (0.2 SD drop in fitness scores, 0.07 SD drop in job performance ratings, 0.8 months less training, and 0.09 fewer promotions; see Table 2). While we see some pretrends among the sample of men on medical-related limited duty status, the pre-match differences are small compared to the post-period changes.

In sum, men who have a recorded medical event that limits their ability to fully engage on the job experience declines in measured job performance, training, and promotion similar to those of mothers after a first birth. “Medical event” as defined here captures a range of physical conditions and limitations, so it is not clear whether we would expect men on a medical status to have larger, smaller, or the same effect sizes as new mothers; the expected size of the impact depends on the nature of men’s injuries and illnesses. Nevertheless, the analysis provides evidence that, like new mothers, men with medical limitations receive lower job performance ratings, score more poorly on assessed physical fitness, and experience promotion slowdowns, suggesting discrimination against mothers specifically does not drive our findings of the impacts of motherhood. That said, only women experience pregnancy and childbirth, meaning that the professional consequences of childbirth—even if similar to those of other physically limiting medical events—will disproportionately effect women in this setting.

4.5 Variation by Maternity Leave Length

Prior to 2015, all DoD branches provided active-duty women 6 weeks of paid maternity leave. In July 2015, the Secretary of the Navy announced that primary caregivers in the Navy and Marine Corps would be entitled to 18 weeks of leave. Women who had given birth earlier in January 2015 or later could retroactively take advantage of the 18-week leave policy before their child turned one. Women who had already returned to work after 6 weeks of leave tended to use the additional 12 weeks of paid time off discontinuously (i.e., as flexible time off). Women who were on leave at the time of the announcement of expanded leave, or gave birth after the announcement, generally took the additional leave consecutively (Bacolod et al., 2022). We analyze these groups separately, referring to the different leave arrangements as “6 weeks + 12 flex” to indicate discontinuous extended leave used as flexible time off, and “18 weeks” to signal the stretch of continuous extended weeks of leave. In early 2016, the Secretary of Defense standardized maternity leave to 12 weeks for all military branches. The 12-week policy applied to pregnancies that began 31 days after the announcement (i.e., pregnancies that began on March 3, 2016 or later, per doctor estimation).

We disaggregate the effects of having a child on women’s outcomes by the length of maternity leave, determined by when she gave birth. The key question of interest is whether longer leave predicts better or worse outcomes when women return to work. 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 into parenthood and a particular leave policy. Selection presents the biggest concern for women who gave birth under the latter part of the 18-week policy and the full 12-week policy, given that these women would have known their leave length was greater than 6 weeks before becoming pregnant. We focus on women who expected to receive 6 weeks of leave at conception and were surprised with the announcement of additional leave. We compare three distinct lengths of leave: 6 weeks, 6 weeks + 12 flexible weeks after returning to work, and 18 weeks. Appendix Table A.6 presents descriptive characteristics of the women who gave birth under these three leave policy groups. There are some differences across groups, though not in any ways that suggest systematic bias in one policy regime or another.[30] We conduct the analysis by defining indicator variables for the "6 weeks + 12 flex" and "18 weeks," with the "6 week" policy as the baseline group. We interact these indicators with the variables in Eq. 2 and make policy-specific predictions for the initial birth drop (i.e., 8 months for physical performance and 1 month for training, education, and promotion), 12 months postbirth, and 24 months postbirth.[31]

Table 3 first replicates the main analysis for the subsample who expected 6 weeks of leave at conception (and their matches). Each panel then shows the results from the regression with policy interactions. Like the full sample, physical performance drops when initially observed 8 months after having a baby across all policy periods. Mothers who had longer maternity leave had larger physical performance declines, particularly those under the "6 weeks + 12 flex" policy. These mothers had returned to work following their initial 6 weeks of leave, then received 12 additional weeks of leave to use by their child's

[30] There are statistically significant differences in percent officers, years of education, and combat job type. If there was selection into fertility, we may expect it to be discontinuous at the policy change. Figure A.4 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. In supplemental analyses we set aside our concern about selection into birth and include mothers who knew they would receive additional leave at conception; that analysis includes more observations in the "18 weeks" period and an additional group of women under the "12 weeks" of leave period. Appendix Table A.7 is the balance table for this sample; the results are in Appendix Table A.8. Because babies born in November–December 2016 could have fallen under either the 18 or 12 week policy depending on date of conception, we exclude these mothers (and their matches) from the policy analysis.

[31] We do not include supervisor-rated job performance or marksmanship. Supervisor ratings for junior enlisted are only available for those who remained in service as of October 2017, which complicates analyses of policy changes that took place in 2015 and 2016. We lack power to subdivide marksmanship scores given sparse observations.

first birthday. An F-test of a differences across policies produces a p-value of 0.054; we note that the standard errors are relatively large given the sample size. Similarly, it is the mothers with flexible leave who had larger gaps in education at $r = 12$ (p(diff), all effects=0.058) and $r = 24$ (p(diff), all effects=0.017), though mothers with flexible leave had the smallest gaps in training by $r = 24$ (p(diff), all effects=0.055). In general, we take this as suggestive evidence that more flexible leave could be detrimental to certain work outcomes.

5 Summary and Conclusions

We use repeated, direct measures of work performance, human capital accumulation, and career advancement to explore the link between the transition to parenthood and workers' outcomes. Our empirical strategy draws on an event study approach based on the timing of a first birth and a matching design that assigns placebo births to observably similar nonparents. We find both men and women's physical performance responds negatively to the transition to parenthood. However, mothers experience large declines in job-related physical performance that persist for two years postbirth, while fathers experience smaller, short-lived declines in job performance that fade by their child's second birthday. Women's supervisor-rated job performance and marksmanship scores also decline in the year after having a child, while men's do not. Mothers' accumulated time in on-the-job training slows, while fathers are largely unaffected. These patterns are consistent with our findings that women's promotion trajectories slow after having a child while men's do not. Among women, promotion delays accumulate over time; the difference in number of promotions between mothers and nonmothers is largest 24 months postbirth. Nonfathers with documented medical limitations at work also experience lower physical performance, supervisor ratings of job performance, training, suggesting it is the medically-induced changes brought about by a birth, rather than discrimination, that affect job performance and promotions.

Gender differences in the promotion effects of parenthood directly lead to pay gaps in this setting. By 24 months postbirth, the average mother would make $40,596 in basic pay according the Marine Corps pay schedule (excluding any bonuses or housing allowances).[32] The impact of birth on promotions means

[32] We use 2022 basic pay scales for this estimate. Basic pay is calculated by years of service and rank. Marines also have

that mothers go from an average of $0 difference in pay compared to matched nonmothers at 10 months before birth to $332 lower pay 24 months postbirth. The female-male wage differential grows from $5,789 10 months before birth to $5,890 at 24 months postbirth (a $101 increase in the wage gap).[33]

Last, additional leave does not improve mothers' work-related outcomes. If the goal of maternity leave is to provide bonding time with children and time for mothers to physically, medically, and mentally recover, this may be good news: job-related motherhood penalties were not exacerbated in this context by more generous parental leave policies, especially in the longer-term, two years after birth. There is some evidence that flexible leave is more disruptive.

Our findings provide a new angle on the longstanding literature that shows parenthood reduces mothers' employment, hours worked, and wages, while having no effect on fathers. We find having a child impacts mothers' job performance and skill accumulation in the first two years of the child's life, highlighting the period after birth as a 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. The present research takes place in an environment with guaranteed health insurance coverage, fully paid parental leave, and (where available) subsidized childcare, but the demands of parenthood still spill into the determinants of wages. Our findings suggest that keeping mothers in the labor force, working the same hours will not eliminate child penalties to women's earnings.

housings allowances that increase with first dependents (thus counteracting the mothers' gap for single mothers) and also increase with rank (thus increasing the gap if promotions are delayed), as well as other bonuses or incentives that may differ for mothers (e.g., combat pay). We focus on basic pay because it aligns most closely to civilian wage and is straightforward to calculate.

[33] Fathers are more advanced in their careers before having a child on average, which generates a prepregnancy gender-based wage differential.

References

Abadie, A. and Spiess, J. (2021). Robust post-matching inference. *Journal of the American Statistical Association*, pages 1–13.

Aguilar-Gomez, S., Arceo-Gomez, E., and De la Cruz Toledo, E. (2019). Inside the black box of child penalties. *Available at SSRN 3497089*.

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

Andresen, M. E. and Nix, E. (2019). What Causes the Child Penalty? Evidence from Same Sex Couples and Policy Reforms. Discussion Papers 902, Statistics Norway, Research Department.

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.

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., Laurita, L., Molloy, M., and Sullivan, R. (2022). Mothers in the military: The effect of maternity policy on leave up-take. *Demography*, 59:787–812.

Bailey, M. J., Sun, S., and Timpe, B. (2021). Prep school for poor kids: The long-run impacts of head start on human capital and economic self-sufficiency. *American Economic Review*, 111(12):3963–4001.

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

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

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.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*.

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.

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

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

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134(3):1405–1454.

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

Cortes, P. and Pan, J. (2020). Children and the remaining gender gaps in the labor market. NBER Working Paper w27980, National Bureau of Economic Research.

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.

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

de Chaisemartin, C. and D'Haultfœuille, X. (2021). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Available at SSRN*.

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 under contract with ICF.

Department of Labor (2022). Most common occupations for women in the labor force. Technical report, U.S. Department of Labor Women's Group, Washington, DC.

Fernández-Kranz, D., Lacuesta, A., and Rodríguez-Planas, N. (2013). The motherhood earnings dip: Evidence from administrative records. *Journal of Human Resources*, 48(1):169–197.

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

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

Healy, O. J. and Heissel, J. A. (2022). Gender disparities in career advancement across the transition to parenthood: Evidence from the marine corps. In *AEA Papers and Proceedings*, volume 112, pages 561–67.

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.

Kim, S. D. and Moser, P. (2021). Women in science: Lessons from the baby boom. Technical Report w29436, National Bureau of Economic Research.

Kleven, H. (2022). The Geography of Child Penalties and Gender Norms: Evidence from the United States. Technical Report w30176, National Bureau of Economic Research, Cambridge, MA.

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. NBER Working Paper w28082, National Bureau of Economic Research.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., and Zweimüller, J. (2019a). Child Penalties Across Countries: Evidence and Explanations. In *AEA Papers and Proceedings*, volume 109, pages 122–126.

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.

Laffers, L. and Schmidpeter, B. (2022). Mothers' Job Search After Childbirth. Technical Report 2113, Johannes Kepler University of Linz.

Lafortune, J., Rothstein, J., and Schanzenbach, D. W. (2018). School finance reform and the distribution of student achievement. *American Economic Journal: Applied Economics*, 10(2):1–26.

Larger, R. B. (2017). Effectiveness of the Marine Corps' Junior Enlisted Performance Evaluation System: An Evaluation of Proficiency and Conduct Marks. Technical report, Naval Postgraduate School, Monterey, CA.

O*NET (2022). Abilities — dynamic strength.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., and Poe, J. (2022). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. Papers, arXiv.org.

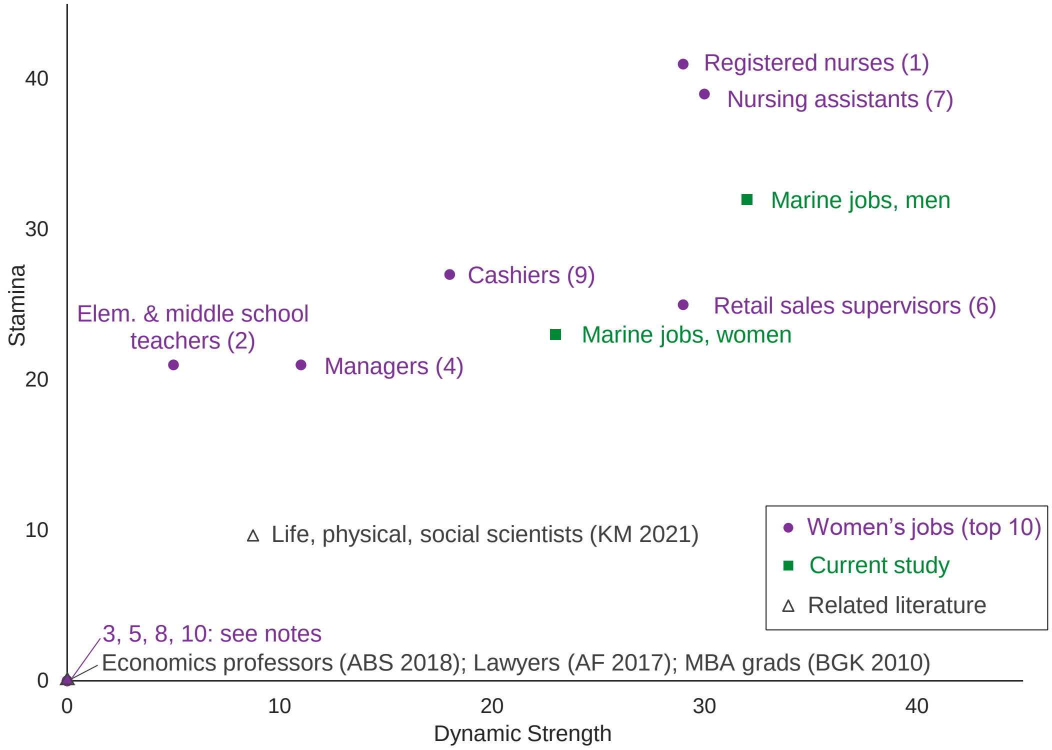
Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

U.S. Marine Corps (2017). Change to Composite Score Point Scale and Classification for CFT and PFT. MARADMIN 084/17, United States Marine Corp.

U.S. Marine Corps (2021). Talent Management 2030.

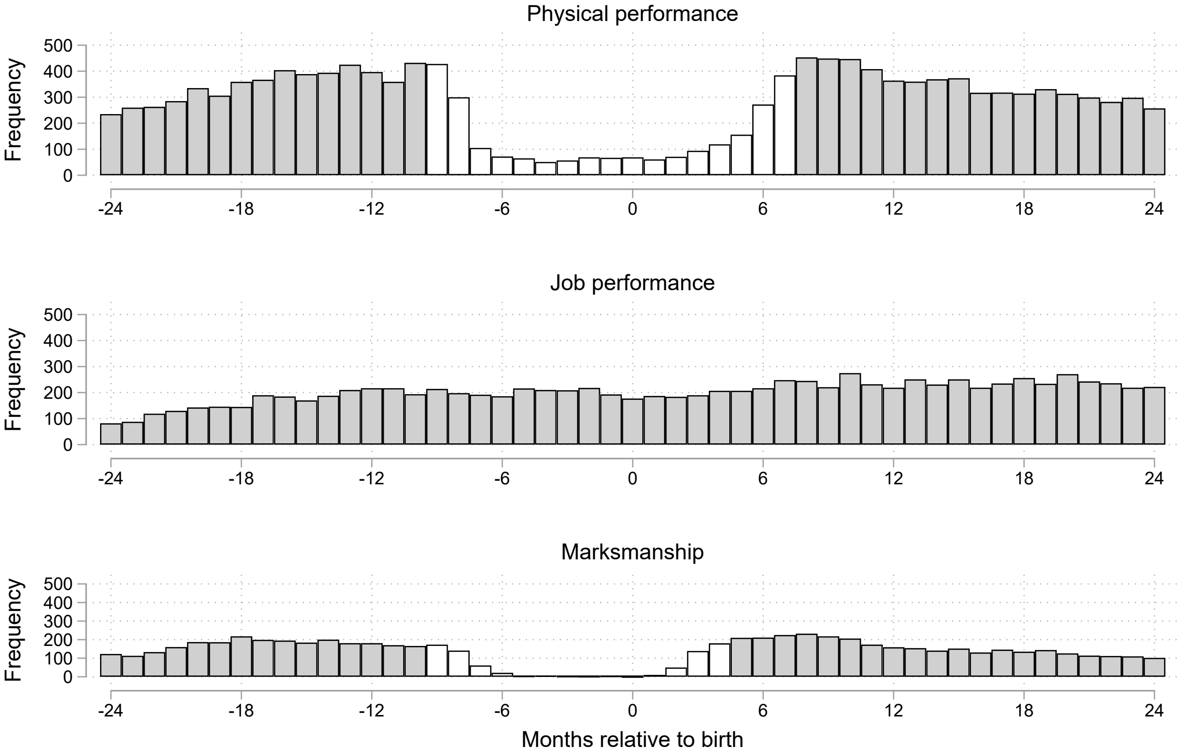
6 Tables and Figures

Figure 1: Distribution of Dynamic Strength and Stamina Required by Various Occupations



Notes: Displays required dynamic strength and stamina based on O*NET data for various jobs (O*NET, 2022). Top 10 occupations for women in the United States come from the United States Department of Labor (Department of Labor, 2022), with the rank given in parentheses. Top women's occupations with no strength or stamina requirements include: administrative assistants (3), customer service representatives (5), accountants and auditors (8), and receptionists (10). Estimates of the job requirements for Marines display the median O*NET score obtained by connecting Marines' military occupations to their civilian equivalents and giving any military occupation without a civilian equivalent the highest observed level (firefighters with dynamic strength = 56 and stamina = 54). Prior related literature has examined changes in productivity and career advancement around birth among physical, biological, and social scientists (Kim and Moser, 2021); economics professors (Antecol et al., 2018); lawyers (Azmat and Ferrer, 2017) and MBA graduates in the corporate and financial sectors (Bertrand et al., 2010). We designate these papers as KM 2021, ABS 2018, AF 2017, and BGK 2010, respectively. Plotted data for the scientist category weights O*NET requirements for occupations in the physical, life and social sciences by their 2019 distribution. Plotted data for the MBA graduate category relies on common occupation destinations for these graduates in the corporate and financial sectors (e.g., chief executive, financial managers, management analysts), which have zero strength or stamina requirements.

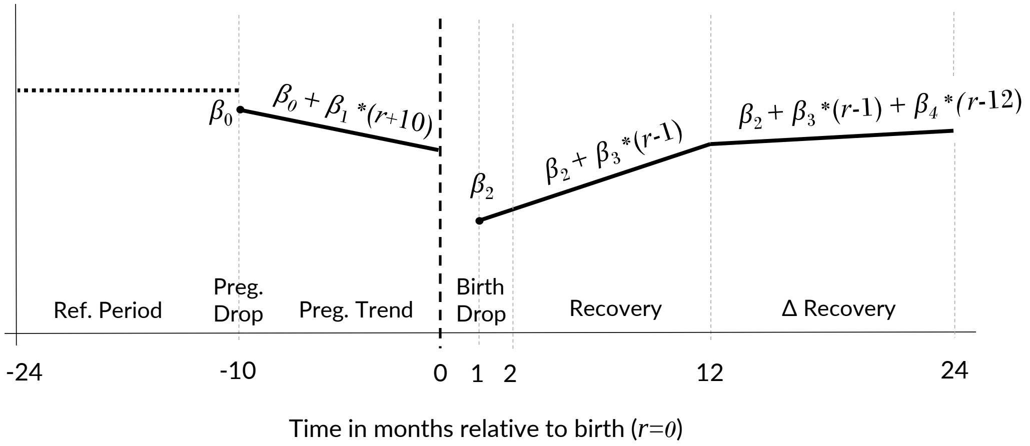
Figure 2: Distribution of Mothers' Outcomes Relative to Birth



Notes: Displays the count of physical performance, job performance, and marksmanship scores for mothers by month relative to birth. Physical performance tests are the most common test among these outcomes; all ranks are expected to take them twice a year (the Physical Fitness Test in January–June and the Combat Fitness Test in July–December). Main analysis excludes physical performance scores at $r = [-9, 7]$ (because mothers did not have to take the tests in pregnancy through 6 months postpartum, and commanders may give them some leeway in month 7) and marksmanship scores at $r = [-9, 4]$ (because mothers did not have to take the tests during pregnancy or while on leave). Semi-dynamic specifications always exclude $r = 0$ due to ambiguity about outcome timing relative to birth. Excluded outcomes are in white.

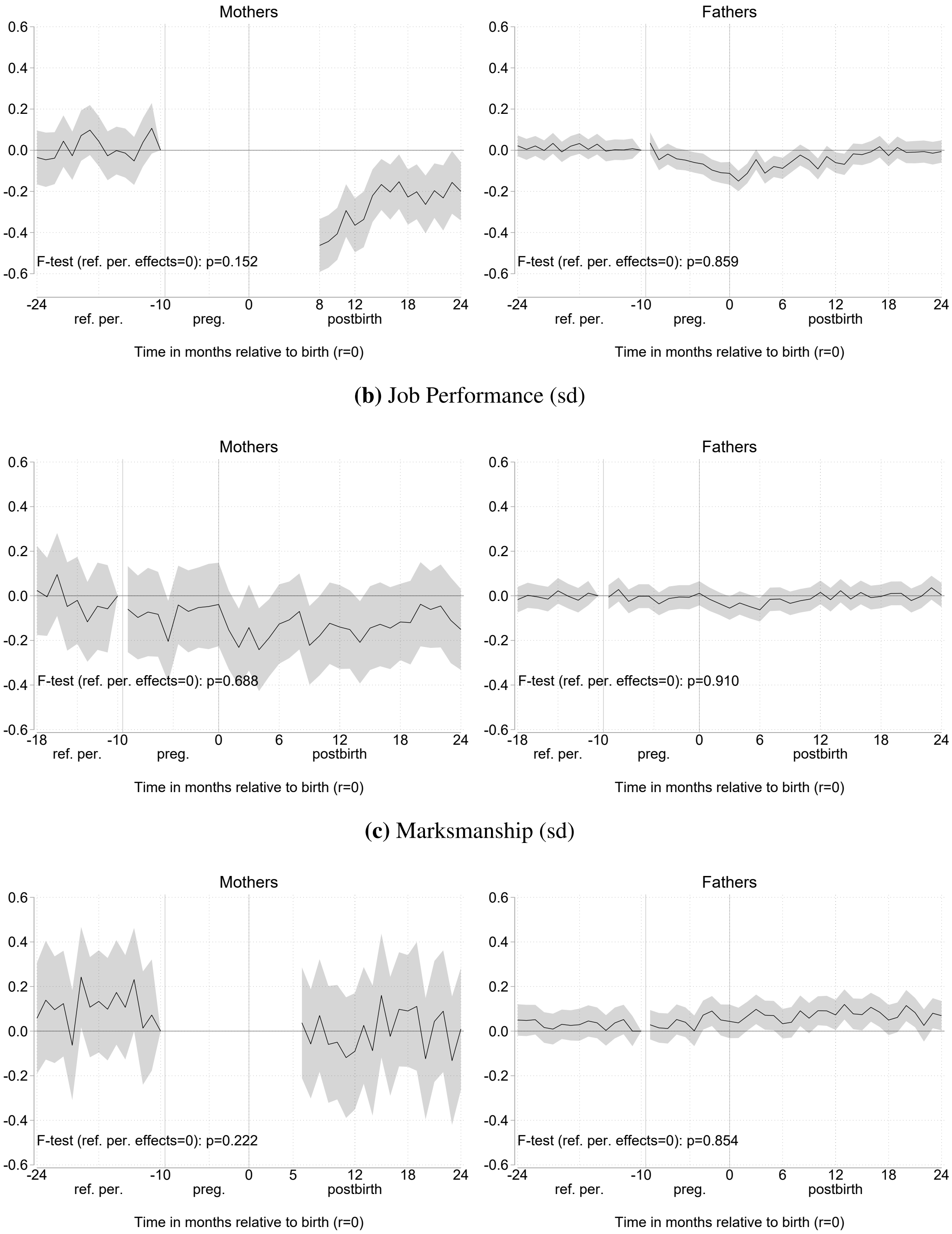
Figure 3: Stylized Representation of Parenthood Effects from the Semi-Dynamic Specification (Eq. 2)

$$Y_{igtr} = \theta_0 PregnancyDrop_{igtr} + \theta_1 PregnancyTrend_{igtr} + \theta_2 BirthDrop_{igtr} + \theta_3 Recovery_{igtr} + \\ \theta_4 \Delta Recovery_{igtr} + \pi P_i + \beta_0(PregnancyDrop_{igtr} \times P_i) + \beta_1(PregnancyTrend_{igtr} \times P_i) + \\ \beta_2(BirthDrop_{igtr} \times P_i) + \beta_3(Recovery_{igtr} \times P_i) + \beta_4(\Delta Recovery_{igtr} \times P_i) + \alpha_g + \phi_t + \varepsilon_{igtr}$$


β_0
$\beta_0 + \beta_1 *(r+10)$
β_2
$\beta_2 + \beta_3 *(r-1)$
$\beta_2 + \beta_3 *(r-1) + \beta_4 *(r-12)$
Ref. Period
Preg. Drop
Preg. Trend
Birth Drop
Recovery
Δ Recovery
-24
-10
0
1
2
12
24
Time in months relative to birth (r=0)

Notes: Figure displays a diagram of parameters defined in Equation 2 where the postbirth drop (β_2) is estimated in the first observed month following pregnancy. For women, we begin measuring the postbirth drop (β_2) for physical fitness performance at 8 months and marksmanship scores at 5 months after birth. We cannot estimate β_0 or β_1, the pregnancy drop and trend, for women's physical fitness outcomes or marksmanship scores because women are not eligible to be assessed when pregnant.

Figure 4: Event Study Estimates of the Impact of Birth on Job Outcomes



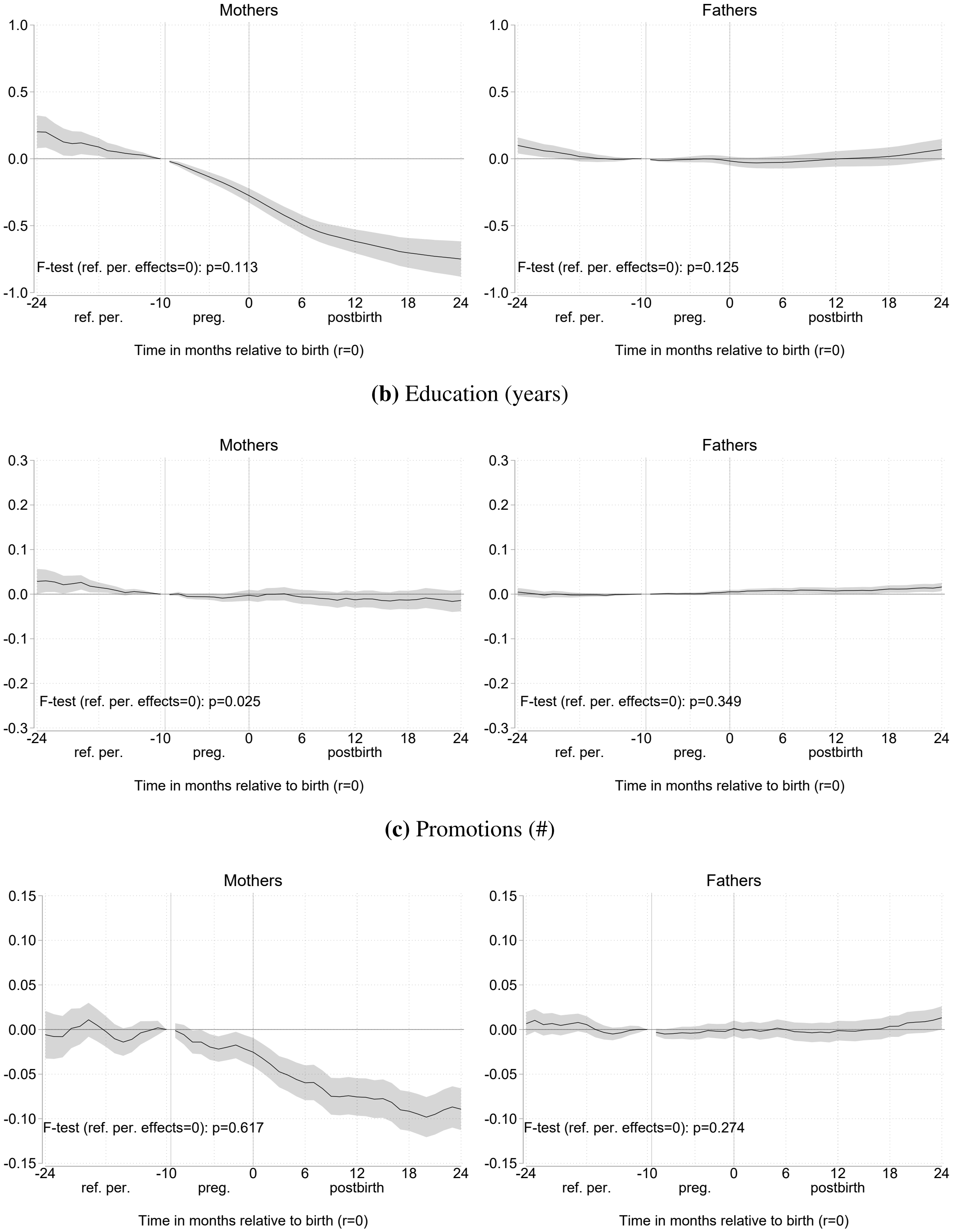
(a) Physical Performance (sd)

(b) Job Performance (sd)

(c) Marksmanship (sd)

Notes: Displays coefficients from event study regressions using the placebo matched sample. Outcomes include standardized scores from (a) physical/combat fitness tests, (b) job performance evaluations, and (c) rifle/pistol marksmanship evaluations. We require nonparents be an exact match on rank, number of months in service, reserve status, and observation year as parents at $r = -10$. Among those, we match parents to a maximum of five most similar nonparents in their propensity to have a child based on age, race/ethnicity, AFQT scores, marital status (and if their spouse is in the military), education level, months of training, occupational field, and most recent physical performance score as of $r = -10$. Regressions include match-group and month-year fixed effects. The reference month is $r = -10$. Vertical lines reflect the start of the pregnancy ($r = -9.5$) and birth ($r = 0$). Standard errors are clustered by individual and match-group and are included as shaded areas representing a 95% confidence interval.

Figure 5: Event Study Estimates of the Impact of Birth on Human Capital and Promotions



(a) Training (months)

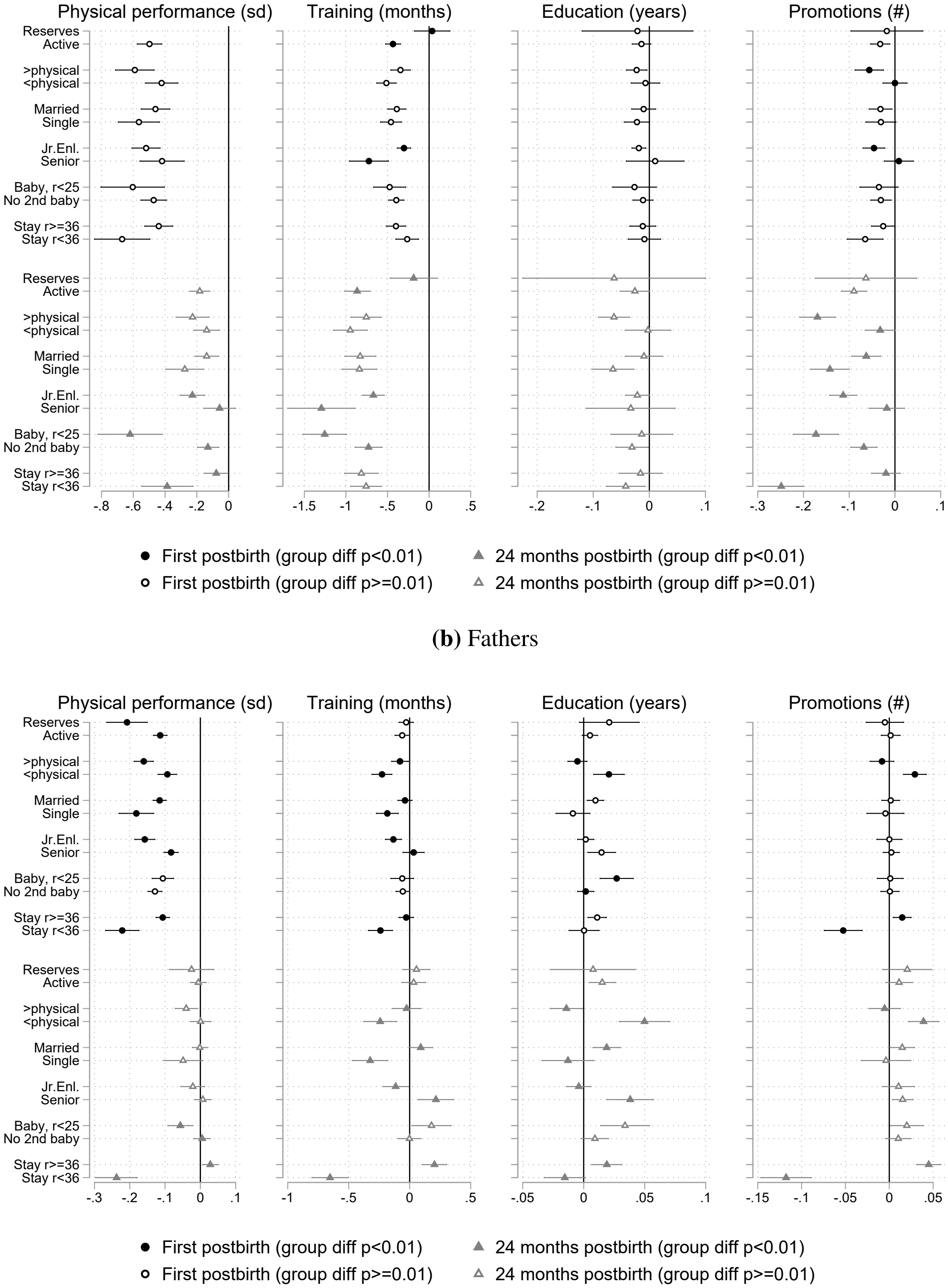
(b) Education (years)

(c) Promotions (#)

Notes: Displays coefficients from event study regressions using the placebo matched sample. Outcomes include (a) cumulative months of training (relative to $r = -10$), (b) cumulative count of years of education (relative to $r = -10$), and (c) cumulative count of promotions (relative to $r = -10$). We require nonparents be an exact match on rank, number of months in service, reserve status, and observation year as parents at $r = -10$. Among those, we match parents to a maximum of five most similar nonparents in their propensity to have a child based on age, race/ethnicity, AFQT scores, marital status (and if their spouse is in the military), education level, months of training, occupational field, and most recent physical performance score as of $r = -10$. Regressions include match-group and month-year fixed effects. The reference month is $r = -10$. Vertical lines reflect the start of the pregnancy ($r = -9.5$) and birth ($r = 0$). Standard errors are clustered by individual and match-group and included as shaded areas representing a 95% confidence interval.

Figure 6: Event Study Estimates of Subgroup Heterogeneity

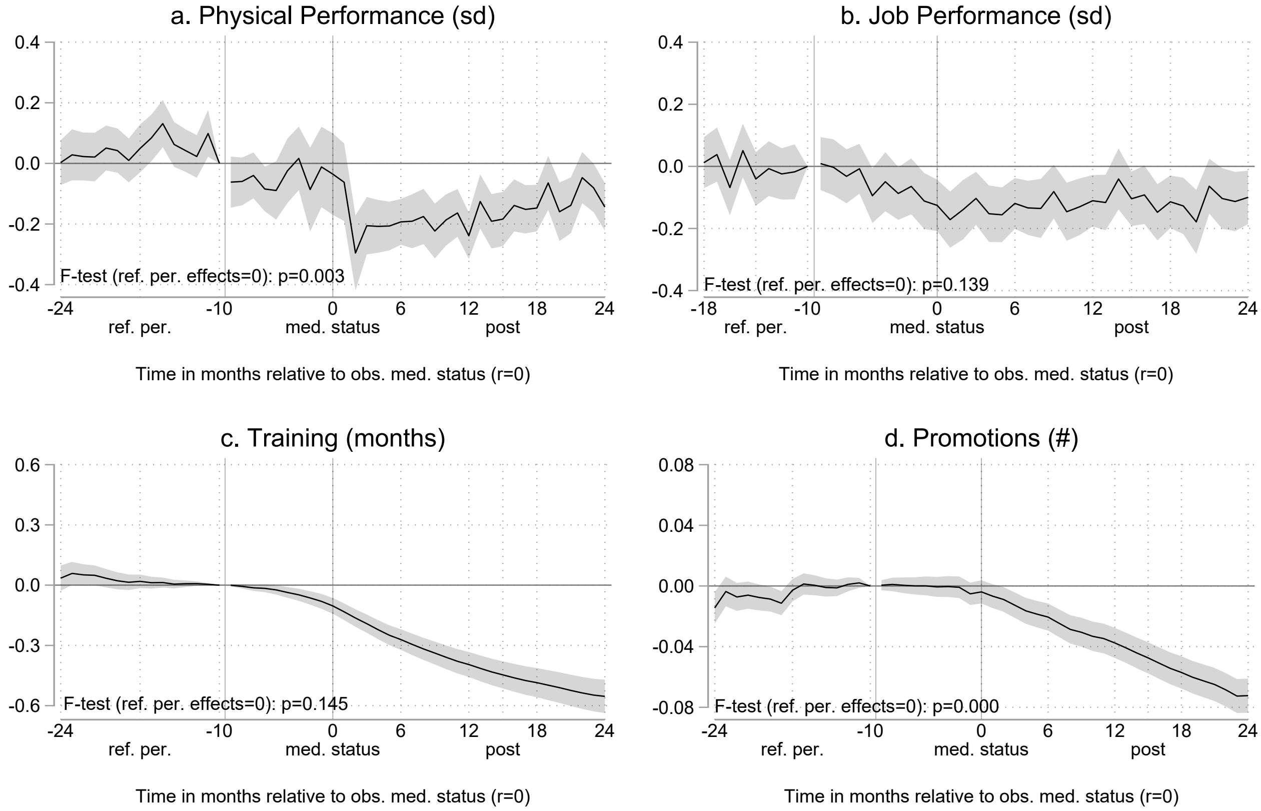
(a) Mothers



(b) Fathers

Notes: Displays gaps in the physical performance, months of training, years of education, and number of promotions relative to prepregnancy between first-time parents and placebo parents across birth events for the first postbirth observation (black line) and 24 month postbirth (light gray line) by subgroups. Each comparison (e.g., reserve vs. active) is based on one regression by interacting an indicator variable with a group indicator (e.g., reserve) with the parameters in Eq. 2. Filled-in markers indicate a statistical significant difference between groups at the $p < 0.01$ level. Classifications for parents are as follows: "Reserves" are not on active duty and likely working a civilian job; "Active" work their military job full-time. ">physical" are those whose military job type above the median physicality level in our sample based on O*NET classification; "<physical" are at or below the median, among those whose jobs are classified by O*NET. "Married" are married at $r = 0$; "Single" are not. "Jr. Enl" are in enlisted grade E1–E4 at $r = -10$; "Senior" are E5 and up or officers. "Baby $r < 25$" have an additional baby within 2 years postbirth; "No 2nd baby" do not. "Stay $r > 36$" stay in the military at least 3 years after the birth event; "Stay $r <= 36$" leave between $r = [24, 36]$. Vertical solid lines reflect a zero effect. Horizontal lies indicate 95% confidence intervals.

Figure 7: Event Study Estimates of the Impact of a Medical Event on Mens' Job Outcomes



Notes: Displays coefficients from event study regressions using assignment to a medical status at $r = 0$ as the event and using a placebo event matched sample. Sample only includes individuals we do not observe having a baby in our study window and who remain in the sample at least $r = [-12, 24]$. Outcomes include standardized scores from (a) physical/combat fitness tests, (b) job performance evaluations, (c) cumulative months of training (relative to $r = -10$), and (d) cumulative count of promotions (relative to $r = -10$). We require non-medical individuals be an exact match on rank, number of months in service, reserve status, and observation year as parents at $r = -10$. Among those, we match individuals with a medical event to a maximum of five most similar non-medical individuals in their propensity to have a medical event based on age, race/ethnicity, AFQT scores, marital status (and if their spouse is in the military), education level, months of training, occupational field, and most recent physical performance score as of $r = -10$. Regressions include match-group and month-year fixed effects. The reference month is $r = -10$. Vertical lines reflect the month after the match ($r = -9.5$) and the start of the observation of medical status ($r = 0$). We include a lag between the match and the start of the observed medical status in case there are delays in updating the medical status. Standard errors are clustered by individual and match-group and are included as shaded areas representing a 95% confidence interval.

Table 1: Descriptive Characteristics

	Women			Men		
Variable	Mothers	Placebos	Difference	Fathers	Placebos	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
	A. Exact match variables					
Months of service	39.209	39.209	0.000	58.570	58.570	0.000
	[42.829]	[42.822]	(0.400)	[49.579]	[49.578]	(0.300)
Officer	0.073	0.073	0.000	0.137	0.137	0.000
	[0.260]	[0.260]	(0.003)	[0.343]	[0.343]	(0.003)
Reservist	0.051	0.051	0.000	0.112	0.112	0.000
	[0.221]	[0.221]	(0.002)	[0.316]	[0.315]	(0.002)
	B. Other matching variables					
Age	22.565	22.700	-0.134*	24.640	24.907	-0.267***
	[4.172]	[4.345]	(0.059)	[4.623]	[4.989]	(0.037)
Black	0.156	0.150	0.006	0.097	0.081	0.016***
	[0.363]	[0.357]	(0.009)	[0.295]	[0.273]	(0.002)
Hispanic	0.224	0.223	0.001	0.143	0.127	0.017***
	[0.417]	[0.416]	(0.010)	[0.350]	[0.333]	(0.003)
Other	0.097	0.091	0.006	0.073	0.081	-0.008**
	[0.296]	[0.288]	(0.008)	[0.260]	[0.272]	(0.003)
Cognitive test (Z-score)	-0.168	-0.181	0.013	0.022	0.130	-0.108***
	[0.938]	[0.927]	(0.019)	[0.997]	[0.991]	(0.010)
Married	0.413	0.393	0.020**	0.672	0.668	0.004
	[0.493]	[0.488]	(0.007)	[0.470]	[0.471]	(0.004)
Military spouse	0.264	0.251	0.013	0.040	0.037	0.003
	[0.441]	[0.434]	(0.008)	[0.195]	[0.188]	(0.002)
Years of education	12.481	12.487	-0.006	12.743	12.761	-0.018
	[1.328]	[1.324]	(0.022)	[1.557]	[1.576]	(0.013)
Recent fitness score	0.068	0.095	-0.027	0.246	0.195	0.051***
	[0.902]	[0.871]	(0.020)	[0.840]	[0.878]	(0.008)
Combat job type	0.048	0.045	0.004	0.288	0.290	-0.002
	[0.214]	[0.206]	(0.005)	[0.453]	[0.454]	(0.004)
Combat support job type	0.626	0.628	-0.002	0.367	0.352	0.016**
	[0.484]	[0.483]	(0.011)	[0.482]	[0.478]	(0.005)
Aviation job type	0.192	0.199	-0.007	0.242	0.250	-0.008
	[0.394]	[0.399]	(0.009)	[0.428]	[0.433]	(0.005)
Avg. analytic weight	1.000	0.211		1.000	0.202	
Observations	2492	12262	14754	24066	120047	144113
Unique individuals	2492	6444	2492	24066	30660	24066

Notes: Displays means (SD in brackets) for parents (Columns 1 and 4) and their respective placebos (columns 2 and 5), and the difference in means (standard error clustered by person and match group in parentheses) between them (Columns 3 and 6) at the time of the match ($r = 10$), weighted by the analytic weight. Required exact match on months of service, rank (e.g., corporal or captain), reservist, and year, with further matching based on predicted propensity score from the remaining variables and their interactions. Includes the average analytic weight, number of unique person-month matches, and number of unique individuals. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 2: Impacts of Childbirth Among First-Time Parents

	Physical	Job Perfor-	Marks-			
	Performance	mance	manship	Training	Education	Promotions
	(sd)	(sd)	(sd)	(months)	(years)	(count)
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Mothers			
1-month effect	–	-0.174***	–	-0.410***	-0.014	-0.032**
		[0.000]		[0.000]	[0.088]	[0.003]
8-month effect	-0.495***	-0.178***	-0.118**	-0.602***	-0.022*	-0.058***
	[0.000]	[0.000]	[0.005]	[0.000]	[0.018]	[0.000]
12-month effect	-0.289***	-0.181***	-0.132**	-0.712***	-0.027*	-0.074***
	[0.000]	[0.000]	[0.006]	[0.000]	[0.012]	[0.000]
24-month effect	-0.184***	-0.071	-0.065	-0.831***	-0.028*	-0.088***
	[0.000]	[0.107]	[0.254]	[0.000]	[0.040]	[0.000]
DV mean (nonparents $r = 24$)	0.14	-0.12	0.21	1.61	12.55	1.31
Unique individuals	8,936	6,489	8,220	8,936	8,936	8,936
Observations	129,495	79,949	55,866	1,155,300	1,152,532	1,155,300
R^2	0.26	0.27	0.19	0.38	0.76	0.79
			B. Fathers			
1-month effect	-0.124***	-0.038*	0.025	-0.057*	0.007*	0.001
	[0.000]	[0.019]	[0.229]	[0.035]	[0.031]	[0.877]
12-month effect	-0.046***	-0.008	0.052**	-0.034	0.008	-0.003
	[0.000]	[0.586]	[0.004]	[0.343]	[0.075]	[0.618]
24-month effect	-0.008	0.030	0.009	0.036	0.014**	0.012
	[0.468]	[0.076]	[0.666]	[0.427]	[0.007]	[0.085]
DV mean (nonparents $r = 24$)	0.02	-0.08	0.25	2.01	12.84	0.96
Unique individuals	54,726	47,188	45,453	54,726	54,726	54,726
Observations	1,865,115	874,845	669,631	12,659,523	12,611,748	12,659,523
R^2	0.25	0.29	0.18	0.40	0.81	0.76

Notes: Displays predicted values from Eq. 2, the semi-parametric specification. Outcomes include (1) standardized (mean=0, SD=1) scores from physical/combat fitness tests conducted 2x per year, (2) standardized scores (mean=0, SD=1) from supervisor-rated job performance evaluations conducted 1-2x per year, (3) standardized scores (mean=0, SD=1) from rifle or pistol tests conducted 1 or fewer times per year, (4) cumulative months of training, (5) cumulative degree counts relative to $r = -10$, and (6) cumulative promotion counts relative to $r = -10$. We exclude women's physical performance scores 9 months before through 7 months after birth because women are not required to take fitness tests during and after pregnancy. We exclude women's marksmanship scores 9 months before through 4 months after birth because women are not required to take marksmanship exams during pregnancy or while on leave. All outcomes for women and men exclude $r = 0$. Regressions include match-group and month-by-year fixed effects. Predicted p-value of whether the value statistically differs from zero are shown in brackets, based on heteroscedasticity-robust F-test and standard errors clustered by match-group and individual. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

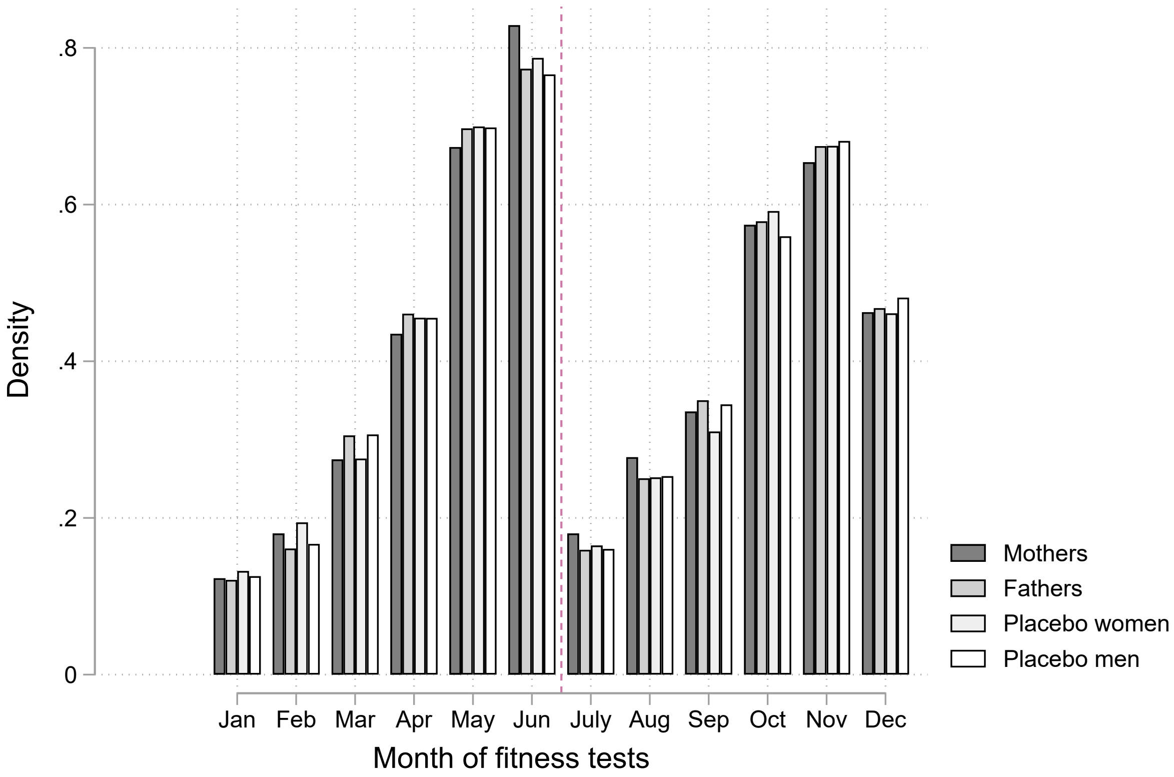
Table 3: Women's Outcomes by Leave Length

	Post-birth drop		12 months post		24 months post		
	Effect size	*p*	Effect size	*p*	Effect size	*p*	N
			A. Physical performance (sd)				
Main effect:	-0.499***	0.000	-0.293***	0.000	-0.138***	0.000	94,796
Effects by paid leave length:							
6 weeks	-0.450***	0.000	-0.295***	0.000	-0.141**	0.002	94,796
6 weeks + 12 flex	-0.795***	0.000	-0.286*	0.011	-0.382**	0.008	
18 weeks	-0.653***	0.000	-0.283***	0.000	-0.042	0.612	
p(diff), all effects	0.054		0.986		0.119		
			B. Training (months)				
Main effect:	-0.331***	0.000	-0.637***	0.000	-0.771***	0.000	858,694
Effects by paid leave length:							
6 weeks	-0.326***	0.000	-0.623***	0.000	-0.709***	0.000	858,694
6 weeks + 12 flex	-0.405	0.050	-0.474	0.051	-0.529	0.087	
18 weeks	-0.302*	0.017	-0.747***	0.000	-1.181***	0.000	
p(diff), all effects	0.909		0.607		0.055		
			C. Years of education				
Main effect:	-0.018	0.073	-0.032*	0.013	-0.028	0.082	856,742
Effects by paid leave length:							
6 weeks	-0.009	0.455	-0.017	0.229	-0.008	0.656	856,742
6 weeks + 12 flex	-0.060*	0.034	-0.139**	0.008	-0.138**	0.006	
18 weeks	-0.042	0.056	-0.053	0.054	-0.076*	0.016	
p(diff), all effects	0.158		0.058		0.017		
			D. Promotions (#)				
Main effect:	-0.037**	0.003	-0.084***	0.000	-0.090***	0.000	858,694
Effects by paid leave length:							
6 weeks	-0.042**	0.003	-0.079***	0.000	-0.083***	0.000	858,694
6 weeks + 12 flex	-0.047	0.301	-0.171***	0.001	-0.151**	0.005	
18 weeks	-0.009	0.735	-0.071*	0.014	-0.095**	0.007	
p(diff), all effects	0.540		0.165		0.468		

Notes: Regressions only include births before March 2016 because mothers could not plan for additional leave announced in July 2015 at conception. Outcomes include physical performance, months of training, years of education, and count of promotion. Postbirth drop is measured at 8 months postbirth for physical performance and at 1 month postbirth for all other outcomes. Regressions include match-group and month-by-year fixed effects. The first row replicates the main analysis for the smaller sample. The next rows display a separate regression from the policy interaction model. "6 weeks" is the predicted mother-placebo gap under the 6-week policy (for babies born December 2014 and prior). "6 weeks + 12 flex" is the predicted mother-placebo gap for mothers who gave birth under the 6-week policy but were retroactively given an additional 12 weeks of leave to use before their baby's first birthday after they had returned to work (for babies born January 2015–mid-May 2015). "18 weeks" is the predicted mother-placebo gap for mothers who gave birth knowing they would have 18 weeks of leave (for babies born mid-May 2015–October 2016). The final row presents the *p*-value for an F-test of whether mother-placebo gaps are the same across all policy periods. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

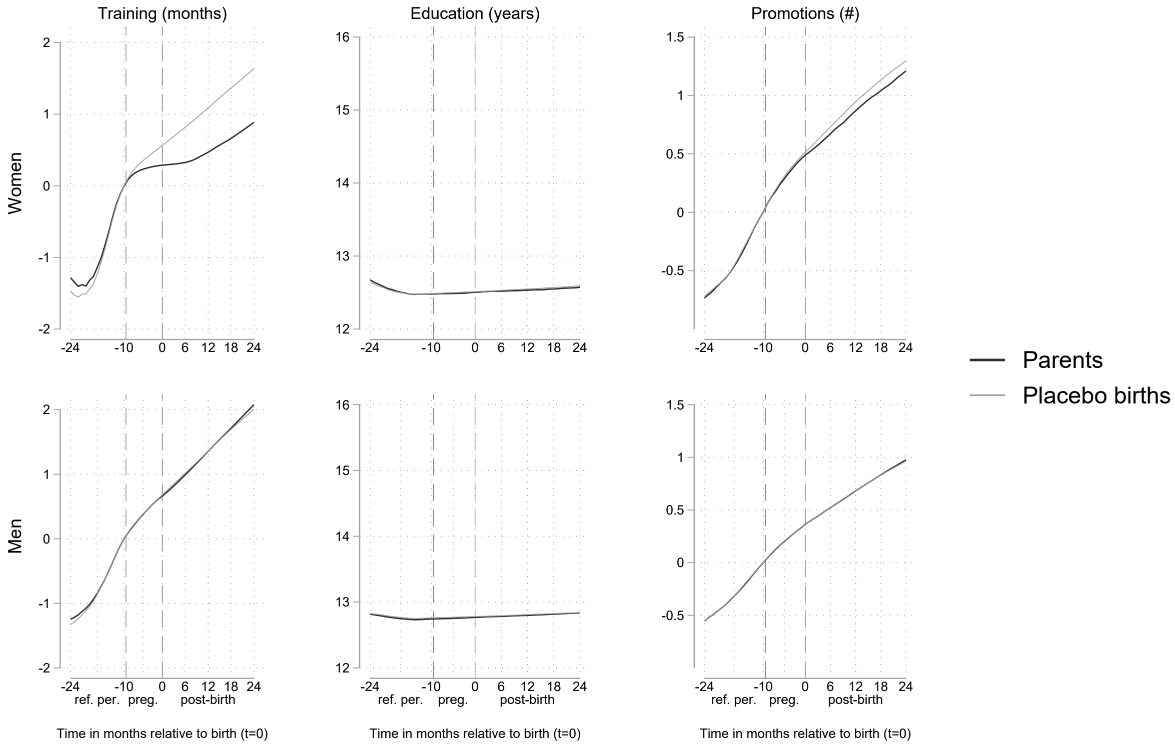
A Supplemental Tables and Figures for Online Publication

Figure A.1: Distribution of physical fitness assessment timing



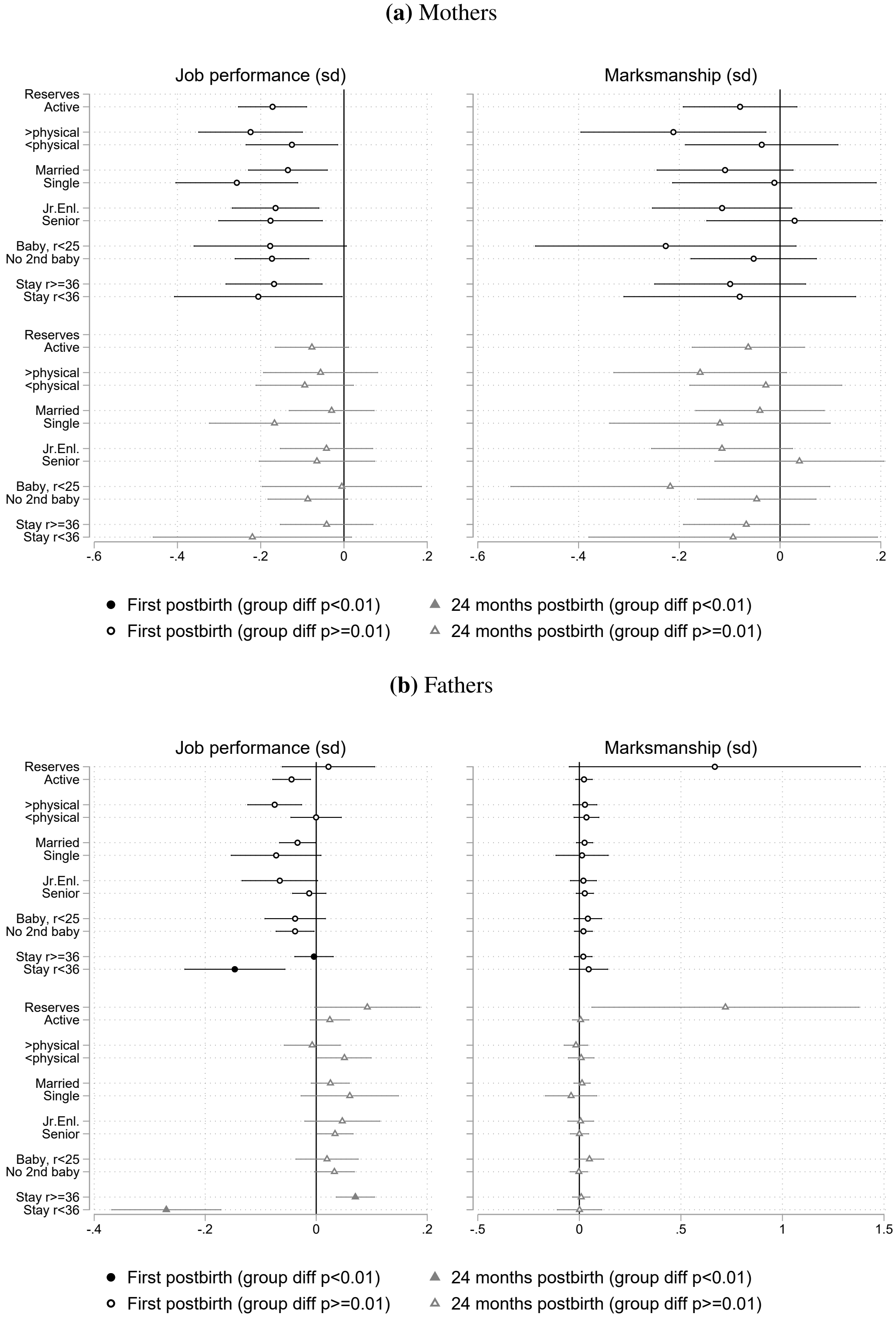
Notes: Displays the distribution of mothers, fathers, placebo women, and placebo men by month of observed fitness assessments in $r = -24, 24]$. The Physical Fitness Test (PFT) test cycle occurs in January–June and the Combat Fitness Test (PFT) test cycle occurs in July–December. Most Marines take the test in the later months of the cycle. Mothers in a unit that scheduled the test in May could run it in June either because they were still under the post-pregnancy fitness waiver in May or because their supervisors allowed them to take it later, even if the mother was off the waiver in May.

Figure A.2: Placebo Birth Estimates of the Impact of Birth on Human Capital and Promotion



Notes: Displays weighted mean levels of cumulative months of training, cumulative count of years of education (relative to $r = -10$), and cumulative count of promotions (relative to $r = -10$) between first-time parents (solid line) and placebo parents (dashed line) over time. Nonparents assigned to placebo births are limited to those whose rank, number of months in service, reserve status, and year is an exact match with parents' 10 months before birth. Among those with an exact match, 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, months of training, occupational field, and most recent physical performance scores.

Figure A.3: Event Study Estimates of Subgroup Heterogeneity in Job Performance & Marksmanship



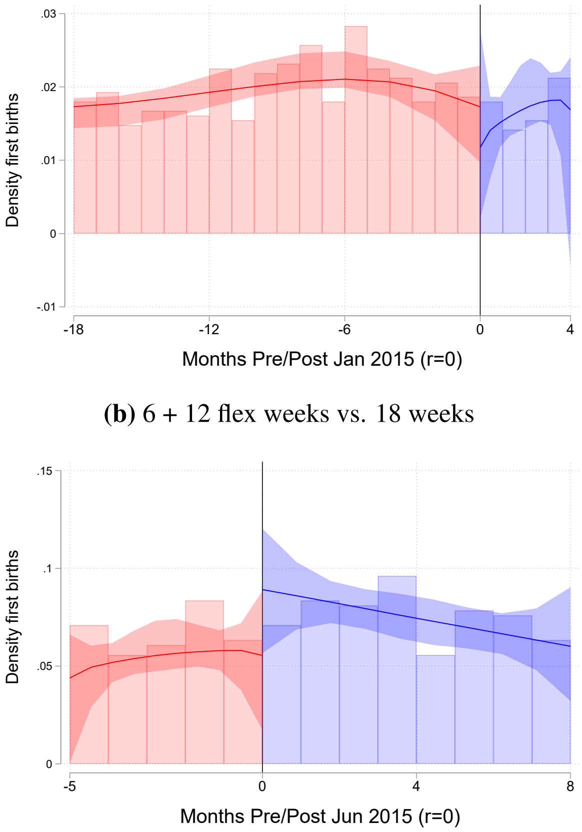
(a) Mothers

(b) Fathers

Notes: Displays gaps in the job performance and marksmanship scores relative to prepregnancy between first-time parents and placebo parents across birth events for the first postbirth observation (black line) and 24 month postbirth (light gray line) by subgroups. Each comparison (e.g., reserve vs. active) is based on one regression by interacting an indicator variable with a group indicator (e.g., reserve) with the parameters in Eq. 2. Filled-in markers indicate a statistical significant difference between groups at the $p < 0.01$ level. Classifications for parents are as follows: "Reserves" are not on active duty and likely working a civilian job; "Active" work their military job full-time. ">physical" are those whose military job type above the median physicality level in our sample based on O*NET classification; "<physical" are at or below the median, among those whose jobs are classified by O*NET. "Married" are married at $r = 0$; "Single" are not. "Jr. Enl" are in enlisted grade E1–E4 at $r = -10$; "Senior" are E5 and up or officers. Vertical solid lines reflect a zero effect. "Baby $r < 25$" have an additional baby within 2 years postbirth; "No 2nd baby" do not. "Stay r >=36" stay in the military at least 3 years after the birth event; "Stay r <36" leave between $r = [24, 36]$. Vertical solid lines indicate a zero effect. Horizontal lies indicate 95% confidence intervals.

Figure A.4: Density of First Births Across Policy Periods

(a) 6 weeks vs. 6 + 12 flex weeks



(b) 6 + 12 flex weeks vs. 18 weeks

Notes: Histogram bars display the density of first births by month before and after r=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 r=0 is not statistically significant in Panel (a) or (b). The sample includes all women in the Marines with a first birth during the time window.

Table A.1: Characteristics of First-Time Parents

	Mothers		Fathers	
	Marines	Civilian	Marines	Civilian
Descriptive characteristics				
Age	23.40	29.97	25.48	31.80
Black	0.15	0.09	0.10	0.07
Hispanic	0.23	0.12	0.14	0.14
Married	0.68	0.81	0.87	0.86
Some college	0.05	0.28	0.05	0.27
College	0.09	0.59	0.16	0.48
Job Classifications				
Mngmt./Business/Science/Arts	0.13	0.57	0.10	0.46
Service	0.07	0.15	0.04	0.11
Sales/Office	0.35	0.24	0.12	0.15
Construction/Maint.	0.18	0.00	0.29	0.15
Production/Moving/Transpo.	0.19	0.03	0.14	0.14
Military	0.08	0.00	0.31	0.00
Military-Specific Characteristics				
Officer	0.07	–	0.14	–
AFQT score (percentile)	58.56	–	63.23	–
GCT score (av=100; sd=20)	103.38	–	111.29	–
N of individuals	2,492	3,638,695	24,059	4,557,719

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 at the month of birth (r=-0). Data on civilians come from the American Community Survey 1-year estimates, 2010 to 2018. 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 Survey. 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.

Table A.2: Coefficients for Impacts of Childbirth Among First-Time Parents

	Physical Performance (sd)	Job Performance (sd)	Marksmanship (sd)	Training (months)	Education (years)	Promotions (count)
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Mothers			
Pregnancy drop	–	-0.067	–	-0.061*	-0.015**	0.003
		(0.053)		(0.029)	(0.005)	(0.008)
Pregnancy trend	–	-0.004	–	-0.028***	-0.001	-0.002*
		(0.008)		(0.003)	(0.001)	(0.001)
Birth drop $(birth - 24mos.)$	-0.495***	-0.174***	-0.106	-0.410***	-0.014	-0.032**
	(0.038)	(0.041)	(0.072)	(0.044)	(0.008)	(0.010)
Recovery $(birth - 24mos.)$	0.051***	-0.001	-0.004	-0.027***	-0.001	-0.004***
	(0.012)	(0.004)	(0.013)	(0.003)	(0.001)	(0.001)
Δ recovery $(13 - 24mos.)$	-0.043**	0.010	0.009	0.018***	0.001	0.003
	(0.014)	(0.007)	(0.018)	(0.003)	(0.001)	(0.002)
DV mean (nonparents $r = 24$)	0.14	-0.12	0.21	1.61	12.55	1.31
Unique individuals	8,936	6,489	8,220	8,936	8,936	8,936
Observations	129,495	79,949	55,866	1,155,300	1,152,532	1,155,300
R^2	0.26	0.27	0.19	0.38	0.76	0.79
			B. Fathers			
Pregnancy drop	0.008	0.021	0.052	-0.035*	-0.001	-0.004
	(0.013)	(0.019)	(0.031)	(0.015)	(0.002)	(0.004)
Pregnancy trend	-0.014***	-0.004	-0.006	0.001	0.000	0.000
	(0.002)	(0.003)	(0.005)	(0.002)	(0.000)	(0.001)
Birth drop $(birth - 24mos.)$	-0.124***	-0.038*	0.025	-0.057*	0.007*	0.001
	(0.009)	(0.016)	(0.021)	(0.027)	(0.003)	(0.005)
Recovery $(birth - 24mos.)$	0.007***	0.003	0.002	0.002	0.000	-0.000
	(0.001)	(0.002)	(0.002)	(0.002)	(0.000)	(0.000)
Δ recovery $(13 - 24mos.)$	-0.004*	0.000	-0.006	0.004*	0.001	0.002*
	(0.002)	(0.003)	(0.004)	(0.002)	(0.000)	(0.001)
DV mean (nonparents $r = 24$)	0.02	-0.08	0.25	2.01	12.84	0.96
Unique individuals	54,726	47,188	45,453	54,726	54,726	54,726
Observations	1,865,115	874,845	669,631	12,659,523	12,611,748	12,659,523
R^2	0.25	0.29	0.18	0.40	0.81	0.76

Notes: Displays coefficients from Eq. 2, the semi-parametric specification. Outcomes include (1) standardized (mean=0, SD=1) scores from physical/combat fitness tests conducted 2x per year, (2) standardized scores (mean=0, SD=1) from supervisor-rated job performance evaluations conducted 1-2x per year, (3) standardized scores (mean=0, SD=1) from rifle or pistol tests conducted 1 or fewer times per year, (4) cumulative months of training, (5) cumulative degree counts relative to $r = -10$, and (6) cumulative promotion counts relative to $r = -10$. We exclude women's physical performance scores 9 months before through 7 months after birth because women are not required to take fitness tests during and after pregnancy. We exclude women's marksmanship scores 9 months before through 4 months after birth because women are not required to take marksmanship exams during pregnancy or while on leave. All outcomes for women and men exclude $r = 0$. Regressions include match-group and month-by-year fixed effects. The parameter "Pregnancy drop" captures any immediate shift from pre-birth to pregnancy, if observed. The parameter "Pregnancy trend" captures trends during pregnancy, if observed. "postbirth drop" is an indicator equal to 1 after birth, starting in $r = 1$ for all men's outcomes; $r = 8$ for women's physical performance; $r = 5$ for women' marksmanship; and $r = 1$ for women's job performance, training, education, and promotion. "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 match group and individual, shown in parentheses. *** $p < 0.001$, ** $p < 0.01$, $^{*}p < 0.05$.

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

	3-Mile Run	Crunches	Pull-Ups	880-Yard-Run	Lifts	Shuttle Run
	(seconds)	(count)	(count)	(seconds)	(count)	(seconds)
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Mothers			
8-month effect	55.026***	-5.027***	-1.243***	10.289***	-3.172***	10.833***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
12-month effect	24.220***	-4.015***	-0.691**	7.129***	-2.626***	7.815***
	[0.000]	[0.000]	[0.004]	[0.000]	[0.000]	[0.000]
24-month effect	19.276**	-2.505***	-0.693**	2.081	-1.540*	1.395
	[0.008]	[0.000]	[0.005]	[0.074]	[0.012]	[0.302]
DV mean (nonparents $r = 24$)	1561.69	98.75	7.10	214.67	69.52	186.16
Unique individuals	8,918	8,925	4,651	8,870	8,871	8,870
Observations	69,950	70,280	20,796	68,542	68,232	68,535
R^2	0.23	0.23	0.31	0.19	0.30	0.20
			B. Fathers			
1-month effect	23.613***	0.072	-0.305***	2.178***	-0.323*	2.021***
	[0.000]	[0.605]	[0.000]	[0.000]	[0.036]	[0.000]
12-month effect	0.610	-0.080	-0.203***	0.697**	0.027	0.883**
	[0.728]	[0.500]	[0.000]	[0.003]	[0.846]	[0.001]
24-month effect	-1.535	0.081	-0.067	-0.306	0.297	0.001
	[0.490]	[0.634]	[0.326]	[0.313]	[0.107]	[0.997]
DV mean (nonparents $r = 24$)	1395.10	102.83	17.16	180.13	102.60	147.87
Unique individuals	54,707	54,716	54,715	54,688	54,693	54,688
Observations	975,367	979,614	966,401	953,194	952,541	953,148
R^2	0.24	0.35	0.23	0.18	0.45	0.20

Notes: Displays coefficients from the semi-parametric specification in Eq. 2 for item-level outcomes by fitness test type. Columns 1–3 show performance on the Physical Fitness Test items, assessed January–June. Limited pull-up outcome data exist for women prior to 2017, during which time they could do push-ups instead. Columns 3–6 show performance on the Combat Fitness Test items, assessed July–December. The 880-yard-run (Column 4) captures scores on the Movement to Contact drill, designed to mimic the stresses of running under pressure in battle. Lifts (Column 5) measure the number of times a Marine can lift a 30-pound ammunition can overhead. Shuttle run (Column 6) displays timed performance on a 300-yard shuttle run obstacle, called the Maneuver Under Fire drill, 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 A.4: Impacts of Childbirth Among First-Time Parents, Controls for Linear Prepregnancy Trends

	Physical Performance (sd)	Job Performance (sd)	Marksmanship (sd)	Training (months)	Education (years)	Promotions (count)
	(1)	(2)	(3)	(4)	(5)	(6)
			A. Mothers			
1-month effect	–	-0.137	–	-0.172**	0.026	-0.033*
		[0.308]		[0.006]	[0.084]	[0.040]
8-month effect	-0.556***	-0.124	-0.096	-0.266**	0.035	-0.060**
	[0.000]	[0.519]	[0.470]	[0.005]	[0.112]	[0.005]
12-month effect	-0.361***	-0.116	-0.107	-0.320**	0.040	-0.075**
	[0.000]	[0.610]	[0.487]	[0.005]	[0.129]	[0.003]
24-month effect	-0.286**	0.024	-0.029	-0.271	0.067	-0.089*
	[0.008]	[0.942]	[0.893]	[0.120]	[0.084]	[0.015]
DV mean (nonparents $r = 24$)	0.14	-0.12	0.21	1.61	12.55	1.31
Unique individuals	8,935	6,489	8,209	8,936	8,936	8,936
Observations	128,863	79,886	55,593	1,150,781	1,148,037	1,150,781
R^2	0.26	0.27	0.19	0.38	0.76	0.79
			B. Fathers			
1-month effect	-0.105***	-0.116*	-0.056	0.070*	0.011*	0.013
	[0.000]	[0.019]	[0.206]	[0.033]	[0.037]	[0.178]
12-month effect	-0.016	-0.144	-0.078	0.175**	0.014	0.019
	[0.581]	[0.087]	[0.239]	[0.003]	[0.140]	[0.204]
24-month effect	0.035	-0.168	-0.176	0.334***	0.023	0.045*
	[0.391]	[0.169]	[0.056]	[0.000]	[0.094]	[0.035]
DV mean (nonparents $r = 24$)	0.02	-0.08	0.25	2.01	12.83	0.96
Unique individuals	54,722	47,184	45,399	54,726	54,726	54,726
Observations	1,847,452	873,331	662,204	12,533,762	12,486,768	12,533,762
R^2	0.25	0.29	0.18	0.40	0.81	0.77

Notes: Displays predicted values from a version of Eq. 2, the semi-parametric specification, that includes a linear slope parameter to control for any prepregnancy trends. All parameters in this model are measured relative to $r = -10$, 10 months before pregnancy, rather than relative to the entire prepregnancy period. See Table 2 for additional notes. *** $p < 0.001$, ** $p < 0.01$, $^{*}p < 0.05$.

Table A.5: Placebo Estimate: Impacts of Placebo Childbirth Among Nonparent Women and Other Matched Nonparent Women

	Physical	Job Perfor-	Marks-			
	Performance	mance	manship	Training	Education	Promotions
	(sd)	(sd)	(sd)	(months)	(years)	(count)
	(1)	(2)	(3)	(4)	(5)	(6)
1-month effect	–	-0.012	–	-0.116	-0.005	0.003
		[0.810]		[0.117]	[0.679]	[0.824]
8-month effect	0.039	-0.008	-0.042	-0.088	-0.006	-0.006
	[0.315]	[0.852]	[0.442]	[0.305]	[0.680]	[0.661]
12-month effect	-0.037	-0.005	0.020	-0.072	-0.006	-0.010
	[0.216]	[0.918]	[0.732]	[0.457]	[0.702]	[0.487]
24-month effect	-0.022	-0.017	-0.053	-0.049	-0.013	0.010
	[0.491]	[0.746]	[0.436]	[0.689]	[0.486]	[0.557]
DV mean (nonparents $r = 24$)	0.05	0.01	0.20	1.88	12.68	1.39
Unique individuals	6,349	4,756	5,668	6,349	6,349	6,349
Observations	78,065	50,374	31,467	690,780	693,097	690,780
R^2	0.25	0.26	0.19	0.40	0.73	0.78

Notes: Placebo test created by removing all mothers from the women's sample, identifying a rank-weighted sample of placebos, matching the placebos to other non-mother women using the LASSO and exact match process, and then running the main analysis. See Table 2 for additional notes. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table A.6: Descriptive Characteristics for the Main Policy Analysis

	Length of Paid Maternity Leave			
	6 wks	6 wks + 12 flex	18 wks	p(diff)
Months of service	38.36	40.17	37.91	72.41
Officer	0.06	0.07	0.11	0.07
Reservist	0.06	0.07	0.04	0.06
Age	22.40	22.77	22.65	25.30
Black	0.16	0.12	0.13	0.15
Hispanic	0.19	0.18	0.24	0.21
Other	0.10	0.12	0.11	0.11
Cognitive test (Z-score)	-0.16	-0.07	-0.18	-0.15
Married	0.40	0.40	0.48	0.75
Military spouse	0.26	0.26	0.30	0.44
Years of Education	12.40	12.60	12.65	12.55
Recent fitness score	0.07	0.00	0.01	-0.12
Combat job type	0.05	0.02	0.08	0.05
Combat support job type	0.63	0.63	0.59	0.62
Aviation job type	0.19	0.21	0.22	0.19
Observations	1423	121	274	1818

Notes: Displays means for mothers by policy period (columns 1–3) and the p-value of an ANOVA test of whether the values differ across groups (column 4 "p(diff)"). Excludes mothers whose first birth occurred March 2016 or later who could have known about extended leave length at the time of conception. Variables are those used in the matching procedure.

Table A.7: Descriptive Characteristics for the Supplementary Policy Analysis

	Length of Paid Maternity Leave				
	6 wks	6 wks + 12 flex	18 wks	12 wks	p(diff)
Months of service	38.36	40.17	40.97	39.43	0.69
Officer	0.06	0.07	0.10	0.07	0.03
Reservist	0.06	0.07	0.04	0.03	0.04
Age	22.40	22.77	23.06	22.42	0.02
Black	0.16	0.12	0.13	0.17	0.18
Hispanic	0.19	0.18	0.26	0.31	0.00
Other	0.10	0.12	0.10	0.07	0.18
Cognitive test (Z-score)	-0.16	-0.07	-0.15	-0.25	0.18
Married	0.40	0.40	0.48	0.38	0.01
Military spouse	0.26	0.26	0.30	0.24	0.17
Years of Education	12.40	12.60	12.67	12.48	0.00
Recent fitness score	0.07	0.00	0.07	0.07	0.88
Combat job type	0.05	0.02	0.07	0.03	0.05
Combat support job type	0.63	0.63	0.59	0.66	0.24
Aviation job type	0.19	0.21	0.21	0.18	0.58
Observations	1423	121	497	393	2434

Notes: Displays means for mothers by policy period (columns 1–4) and the p-value of an ANOVA test of whether the values differ across groups (column 5 "p(diff)"). Excludes mothers whose first birth was in November–December 2016 due to ambiguity about the policy for such mothers. Variables are those used in the matching procedure.

Table A.8: Supplementary Analysis: Women's Outcomes by All Maternity Leave Lengths

	Post-birth drop		12 months post		24 months post		
	Effect size	*p*	Effect size	*p*	Effect size	*p*	N
			A. Physical performance (sd)				
Main effect:	-0.495***	0.000	-0.286***	0.000	-0.183***	0.000	126,513
Effects by paid leave length:							
6 weeks	-0.453***	0.000	-0.296***	0.000	-0.139**	0.002	126,513
6 weeks + 12 flex	-0.827***	0.000	-0.276*	0.014	-0.373*	0.010	
18 weeks	-0.582***	0.000	-0.278***	0.000	-0.159**	0.009	
12 weeks	-0.419***	0.000	-0.267***	0.000	-0.318***	0.000	
p(diff), all effects	0.081		0.979		0.127		
			B. Training (months)				
Main effect:	-0.401***	0.000	-0.695***	0.000	-0.814***	0.000	1,130,047
Effects by paid leave length:							
6 weeks	-0.328***	0.000	-0.624***	0.000	-0.713***	0.000	1,130,047
6 weeks + 12 flex	-0.406*	0.049	-0.483*	0.046	-0.549	0.076	
18 weeks	-0.420***	0.000	-0.795***	0.000	-1.073***	0.000	
12 weeks	-0.634***	0.000	-0.887***	0.000	-0.928***	0.000	
p(diff), all effects	0.119		0.244		0.164		
			C. Years of education				
Main effect:	-0.014	0.104	-0.026*	0.016	-0.026	0.052	1,127,373
Effects by paid leave length:							
6 weeks	-0.009	0.454	-0.017	0.228	-0.008	0.669	1,127,373
6 weeks + 12 flex	-0.060*	0.032	-0.139**	0.008	-0.136**	0.006	
18 weeks	-0.007	0.707	-0.016	0.479	-0.045	0.088	
12 weeks	-0.021	0.179	-0.029	0.163	-0.032	0.285	
p(diff), all effects	0.348		0.153		0.095		
			D. Promotions (#)				
Main effect:	-0.030**	0.005	-0.074***	0.000	-0.089***	0.000	1,130,047
Effects by paid leave length:							
6 weeks	-0.043**	0.003	-0.078***	0.000	-0.084***	0.000	1,130,047
6 weeks + 12 flex	-0.049	0.284	-0.168***	0.001	-0.158**	0.003	
18 weeks	-0.020	0.345	-0.082***	0.001	-0.090***	0.001	
12 weeks	0.006	0.805	-0.021	0.444	-0.087**	0.004	
p(diff), all effects	0.315		0.058		0.618		

Notes: Regressions exclude birth after March 2016 births given that these women would have known about extended leave before becoming pregnant.Outcomes include physical performance, months of training, years of education, and count of promotion. Postbirth drop is measured at 8 months postbirth for physical performance and at 1 month postbirth for all other outcomes. Regressions include match-group and month-by-year fixed effects. The first row replicates the main analysis for the smaller sample. The next rows display a separate regression from the policy interaction model. "6 weeks" is the predicted mother-placebo gap under the 6-week policy (for babies born December 2014 and prior). "6 weeks + 12 flex" is the predicted mother-placebo gap for mothers who gave birth under the 6-week policy but were retroactively given an additional 12 weeks of leave to use before their baby's first birthday after they had returned to work (for babies born January 2015–mid-May 2015). "18 weeks" values show the predicted mother-placebo difference for mothers who gave birth when 18 weeks of leave was in place but who did not know of this change at the time of conception. For this policy 12 weeks of the leave could be used flexibly before the baby's first birthday (for babies born mid-May 2015–February 2016). "12 weeks" is the predicted mother-placebo gap for mothers who gave birth knowing they would have 12 weeks of leave to use immediately following birth (for babies born January 2017 and later). The final row presents the *p*-value for an *F*-test of whether mother-placebo differences are the same across all policy periods. *** $p < 0.001$, ** $p < 0.01$, $^{*}p < 0.05$.

B. Data Appendix for Online Publication

Our data are from U.S. Department of Defense administrative records and cannot be shared. Below, we review several alternative specifications to our preferred model.

B.1 Alternative Specifications

Our preferred empirical strategy prioritizes identifying an appropriate set of comparison cases to model counterfactual trends that first-time parents would have experienced absent a birth. Tables B.1 and B.2 explore several alternative models for mothers and fathers, respectively. The table columns are our main outcomes of interest, while each row in a given segment shows results for alternative specifications.

Row (1) is a standard two-way fixed effect (TWFE) event study. The comparison group is all same-gender Marines who do not have a baby and remain in the Marine Corps at least three years. This model will involve 2 × 2 comparisons where the counterfactual draws on already treated units, which can lead to what we term TWFE bias. In our setting, we are less worried about TWFE bias because we have a large number of "never-treated" comparison individuals, meaning that most of our overall estimate of counterfactual time trends will come from comparisons of untreated-to-treated against never-treated controls. The main concern with this model is that parents are not the same as average nonparents and thus nonparents do not provide a helpful counterfactual. We cannot use the TWFE approach to estimate training or promotion impacts because we measure these outcomes cumulatively, and cumulative measures requires a pregnancy starting point that the comparison group (not assigned a placebo birth) does not have. We specifically run the following model:

$$\begin{aligned} Y_{it} =& \alpha_i + \phi_t + \beta_0 PregnancyDrop_{it} + \beta_1 PregnancyTrend_{it} + \beta_2 BirthDrop_{it} \\ & + \beta_3 Recovery_{it} + \beta_4 \Delta Recovery_{it} + \varepsilon_{it} \end{aligned} \tag{B.1}$$

Here, β_0 captures any immediate intercept shift and 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 regression excludes $r = 0$ due to ambiguity about the timing of the outcome relative to birth for all outcomes; it also excludes fitness and marksmanship scores for excluded months. Coefficient β_2 represents the acute postnatal birth drop (if any) in the outcome in the first 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 second year of life ($t = [13, 24]$). All parameters are measured relative to the prepregnancy average ($t \leqslant -10$).

Row (2) is a stacked TWFE model, which identifies cohorts of units treated at the same time, excludes any already-treated units from each cohort, stacks the cohorts, and then runs TWFE models across cohort-specific groups.[34] Each parent is connected to five nearest neighbor nonparents from the LASSO prediction model such that parents' births and nonparents' placebo births occur in the same month and year, which eliminates the negative weighting that can occur in traditional TWFE models (Callaway and Sant'Anna, 2021; Cengiz et al., 2019; Sun and Abraham, 2021). The LASSO model used to match parents and nonparents is the same as in the preferred analysis and includes a linear control for months of service, an indicator for officer, and an indicator for active/reserve status. However, we do not require an exact match on months of service, exact rank, or active/reserve status. Once we have the treated and comparison groups, we run the following model:

$$\begin{aligned} Y_{igt} =& \alpha_g + \theta_0 PregnancyDrop_{igt}^{all} + \theta_1 PregnancyTrend_{igt}^{all} + \theta_2 BirthDrop_{igt}^{all} + \\ & \theta_3 Recovery_{igt}^{all} + \theta_4 \Delta Recovery_{igt}{}^{parents} + \beta_0 PregnancyDrop_{igt}^{parents} + \\ & \beta_1 PregnancyTrend_{igt}^{parents} + \beta_2 BirthDrop_{igt}^{parents} + \beta_3 Recovery_{igt}^{parents} + \\ & \beta_4 \Delta Recovery_{igt}{}^{parents} + \varepsilon_{igt} \end{aligned} \tag{B.2}$$

The key difference from Eq. 2 is that it excludes ϕ_t, the month-year fixed effect, because the time relative to birth and month-year is exactly the same within match groups. It also does not contain binned endpoints, because we only include $r = [-24, 24]$. Otherwise, the interpretation is the same as the main model. The potential problem for this model is that the treated and control group might differ on characteristics we know determine the outcome (i.e., months of service, rank, and active/reserve status).

[34] In a stacked approach, regression estimates from each treatment-time cohort are combined using variance weighting to recover a single estimate of the impact across cohorts. Recently proposed alternative estimators, for example by Callaway and Sant'Anna (2021) and Sun and Abraham (2021), use other approaches to weighting each cohort-specific treatment estimate.

Row (3) is similar to our preferred model, but it only includes data from $r = [-24, 24]$ and removes the binned endpoints at $r = -25$ and $r = 25$. The binned endpoints were necessary for modeling both time relative to the match and month-year fixed effects. The model in Row (3) removes the month-year fixed effects, which requires the assumption that the relative time trends in Eq. 2 sufficiently capture counterfactual time trends. We use the same set of individuals as in the preferred specification. The model is identical to the Stacked TWFE model; the difference is that the stacked TWFE model exact matched on month-year while Row (3) exact matches on months of service, rank, active/reserve status, and year.

Row (4) is our preferred model provided for reference. It includes binned endpoints, month-year fixed effects, and exact matching on months of service, rank, active/reserve status, and calendar year. It is synonymous with Eq. 2.

B.2 Alternative Specification Results

The top panels of Tables B.1 and B.2 show a series of F-tests assessing the pretrends in $r = [-24, -10]$ by outcome for the different models. This is analagous to the F-test displayed in Figures 4 and 5. If the untreated group is a good counterfactual, we expect these estimates to be zero. There is some evidence of pretrends for both placebo estimates in education, with a p-value of 0.030 for the model with no time fixed effects and 0.025 for the preferred model with time fixed effects. Given the number of outcomes we examine, this could happen by chance, but for this reason we take the women's education outcomes with a grain of salt. The other exact match pretrends are null for the mothers. The standard TWFE models is more precisely significant for the pretrends in education ($p = 0.000$), while the stacked TWFE model is significant for physical performance ($p = 0.009$), marskmanship ($p = 0.023$), and training ($p = 0.000$). The fathers show some evidence of pretrends in 25% of the standard TWFE outcomes, 67% of the stacked TWFE outcomes, and none of the two exact-matched placebo outcomes. This highlights the importance of exact matching in our setting.

The middle and bottom panels of the tables show the predicted value for the given outcome at $r = 12$

and $r = 24$, respectively. The broad takeaways are generally consistent across all models for women: there are large drops in physical performance that never return to pre-pregnancy levels, while training, education, and promotions remain below expectations at $r = 24$. However, the size of these predictions differ by model and highlight the importance of choosing the best comparison group. The parallel trends assumption means that the nonparents must represent a good counterfactual to the parents in the postperiod. Parallel pretrends offers support for this assumption, but parallel pretrends do not guarantee parallel postreeatment counterfactual trends. In our case, parents' (unobserved) counterfactual postbirth trajectory may differ from nonparents' (observed) postbirth trajectory. As an example, women in combat roles may have better expected physical fitness trajectories in the long-run but are also less likely to become mothers. These women in combat jobs would not be a good counterfactual to the average mothers. For this reason, we prefer the conditional parallel trends assumption required in the exact matching strategies. Both exact-match placebo birth strategies (with and without time fixed effects) produce almost identical results.

We know that promotion is mechanically tied to rank and time in service, so it is particularly important to ensure parents and nonparents match on these characteristics in the preperiod. We would not want to match a low-ranking officer to mid-ranking enlisted, even if they have similar rates of promotion in the preperiod, as their subsequent expected promotion trajectories differ even in the absence of a child. Indeed, when comparing the stacked fixed effect model to the preferred exact-match model, the promotion gap is 55% larger for mothers and, for fathers, flips direction and becomes statistically significant.

Our approach that prioritizes defining an appropriate counterfactual while still considering how to minimize TWFE bias may be useful in cases where cells sizes are too small to isolate distinct cohorts of cases treated at the same time and maintain a sufficient number of comparison cases, especially in settings where other variables (e.g., months of service) are particularly important to consider for creating a counterfactual group. For instance, with yearly state-level data where total observations are limited by the total number of U.S. states, researchers might consider grouping proximate years rather than defining groups of states with the same precise treatment year.

Table B.1: Specification checks for alternative approaches for women

	Physical Performance (sd)	Job Performance (sd)	Marksmanship (sd)	Training (months)	Education (years)	Promotions (count)
	(1)	(2)	(3)	(4)	(5)	(6)
			A. F-test (prepregnancy effects=0), p-value			
Standard TWFE	0.352	0.736	0.127	–	0.000	–
Stacked TWFE	0.009	0.669	0.023	0.000	0.134	0.210
Placebo event, no time FE	0.096	0.778	0.256	0.148	0.031	0.683
Placebo event, TWFE (preferred)	0.152	0.794	0.222	0.113	0.025	0.617
			B. 12-month effect			
Standard TWFE	-0.186***	-0.070	-0.087*	–	-0.100***	–
	[0.000]	[0.052]	[0.041]	–	[0.000]	–
Stacked TWFE	-0.290***	-0.083*	-0.123*	-0.775***	-0.023	-0.105***
	[0.000]	[0.045]	[0.013]	[0.000]	[0.073]	[0.000]
Placebo event, no time FE	-0.290***	-0.171***	-0.117*	-0.711***	-0.026*	-0.073***
	[0.000]	[0.000]	[0.017]	[0.000]	[0.013]	[0.000]
Placebo event, TWFE (preferred)	-0.289***	-0.181***	-0.132**	-0.712***	-0.027*	-0.074***
	[0.000]	[0.000]	[0.006]	[0.000]	[0.012]	[0.000]
			C. 24-month effect			
Standard TWFE	-0.129***	-0.178***	-0.027	–	-0.130***	–
	[0.000]	[0.000]	[0.605]	–	[0.000]	–
Stacked TWFE	-0.174***	-0.015	-0.018	-0.895***	-0.023	-0.136***
	[0.000]	[0.756]	[0.758]	[0.000]	[0.148]	[0.000]
Placebo event, no time FE	-0.179***	-0.052	-0.060	-0.829***	-0.027*	-0.088***
	[0.000]	[0.246]	[0.298]	[0.000]	[0.042]	[0.000]
Placebo event, TWFE (preferred)	-0.184***	-0.071	-0.065	-0.831***	-0.028*	-0.088***
	[0.000]	[0.107]	[0.254]	[0.000]	[0.040]	[0.000]

Notes: Tables displays tests and predicted outcomes for alternative specifications for various outcomes. Standard TWFE model is a traditional event study where the comparison group is all same-gender Marines who do not have a baby and remain in the Marine Corps for at least three years; the comparison units are not matched at a particular point in time and do not have estimates for time relative to a placebo birth. Stacked TWFE model uses exact year-month matching with the five nearest neighbors, where time relative to birth and calendar date are synonymous within groups. This model does not exact match on months of service, rank, or active/reserve status; the underlying matching within year-month does include these variables and is identical to the preferred matching model. The model includes estimates for the placebos, then tests whether the parents differ from those patterns. Exact Match, no time FE model is the same matching process as the preferred model, but does only includes $r = [-24, 24]$ rather than binning $r < -24$ and $r > 24$. The model does not include exact month-year fixed effects. The model includes estimates for the placebos, then tests whether the parents differ from those patterns. Exact Match, time FE model is the preferred model that bins $r < -24$ and $r > 24$. The model includes exact month-year fixed effects. The model includes estimates for the placebos, then tests whether the parents differ from those patterns. Details included in Data Appendix. Training and promotion outcomes excluded from the standard TWFE model because it requires a starting point for the count; parents' count starts at $r = -10$. The first panel tests for pretrends with the p-value of an F-test of whether the points estimates for $r = [-24, -11]$ statistically differ from zero. The second and third panel predicts the effect for parents at $r = 12$ and $r = 24$, respectively. *** $p < 0.001$, ** $p < 0.01$, $^{*}p < 0.05$.

Table B.2: Specification checks for alternative approaches for men

	Physical Performance (sd)	Job Performance (sd)	Marksmanship (sd)	Training (months)	Education (years)	Promotions (count)
	(1)	(2)	(3)	(4)	(5)	(6)
A. F-test (prepregnancy effects=0), p-value						
Standard TWFE	0.804	0.323	0.988	–	0.000	–
Stacked TWFE	0.000	0.949	0.988	0.000	0.000	0.000
Placebo event, no time FE	0.819	0.119	0.466	0.252	0.332	0.399
Placebo event, TWFE (preferred)	0.859	0.163	0.355	0.125	0.349	0.274
B. 12-month effect						
Standard TWFE	0.011*	0.059***	0.058***	–	-0.018***	–
	[0.029]	[0.000]	[0.000]	–	[0.000]	–
Stacked TWFE	-0.033***	0.022	0.048***	-0.035	0.015***	-0.043***
	[0.000]	[0.061]	[0.000]	[0.282]	[0.000]	[0.000]
Placebo event, no time FE	-0.045***	-0.003	0.060**	-0.032	0.008	-0.002
	[0.000]	[0.844]	[0.001]	[0.371]	[0.073]	[0.742]
Placebo event, TWFE (preferred)	-0.046***	-0.008	0.052**	-0.034	0.008	-0.003
	[0.000]	[0.586]	[0.004]	[0.343]	[0.075]	[0.618]
C. 24-month effect						
Standard TWFE	-0.002	0.007	0.040**	–	-0.018***	–
	[0.756]	[0.543]	[0.004]	–	[0.000]	–
Stacked TWFE	-0.000	0.074***	0.045**	0.057	0.025***	-0.046***
	[0.978]	[0.000]	[0.005]	[0.157]	[0.000]	[0.000]
Placebo event, no time FE	-0.005	0.041*	0.019	0.038	0.015**	0.013
	[0.643]	[0.016]	[0.352]	[0.399]	[0.006]	[0.070]
Placebo event, TWFE (preferred)	-0.008	0.030	0.009	0.036	0.014**	0.012
	[0.468]	[0.076]	[0.666]	[0.427]	[0.007]	[0.085]

Notes: Tables displays tests and predicted outcomes for alternative specifications for various outcomes. Standard TWFE model is a traditional event study where the comparison group is all same-gender Marines who do not have a baby and remain in the Marine Corps for at least three years; the comparison units are not matched at a particular point in time and do not have estimates for time relative to a placebo birth. Stacked TWFE model uses exact year-month matching with the five nearest neighbors, where time relative to birth and calendar date are synonymous within groups. This model does not exact match on months of service, rank, or active/reserve status; the underlying matching within year-month does include these variables and is identical to the preferred matching model. The model includes estimates for the placebos, then tests whether the parents differ from those patterns. Exact Match, no time FE model is the same matching process as the preferred model, but does only includes $r = [-24, 24]$ rather than binning $r < -24$ and $r > 24$. The model does not include exact month-year fixed effects. The model includes estimates for the placebos, then tests whether the parents differ from those patterns. Exact Match, time FE model is the preferred model that bins $r < -24$ and $r > 24$. The model includes exact month-year fixed effects. The model includes estimates for the placebos, then tests whether the parents differ from those patterns. Details included in Data Appendix. Training and promotion outcomes excluded from the standard TWFE model because it requires a starting point for the count; parents' count starts at $r = -10$. The first panel tests for pretrends with the p-value of an F-test of whether the points estimates for $r = [-24, -11]$ statistically differ from zero. The second and third panel predicts the effect for parents at $r = 12$ and $r = 24$, respectively. $^{***}\ p < 0.001$, $^{**}\ p < 0.01$, $^{*}p < 0.05$.

C. Promotion Simulation

Our three measures of job performance directly contribute to the promotion points calculation. In this appendix, we simulate how physical fitness, supervisor ratings of job performance, and rifle marksmanship contribute to the promotion gap for someone who last promoted at $r = -10$. Larger (2017) and U.S. Marine Corps (2017) provide additional details on the promotion point calculation details. Here, we convert the raw points in the data into their promotion point equivalent and simulate changes in promotion points according to the calculation rules.

In our main analysis, the effect as measured in standard deviations is largest in physical fitness, but the promotion points system weights job performance more than physical fitness. PFT, CFT, and rifle scores produce a maximum of 166.7 points each (500 points total; 333.3 points total for fitness). Job performance points use the average of the ProCon evaluations at a given rank (with a maximum of 1,000 points). For an E4, the median (interquartile range) for points are 157 (IQR= $140 - 163$) for rifle, 320 (IQR= $287 - 326$) for combined fitness, and 890 (IQR= $860 - 910$) for cumulative ProCons. There are a maximum of 100 points from self-education and 200 points from various bonuses (e.g., recruitment).

At $r = 8$, the first time we observe all three outcomes, mothers who took either the PFT or CFT would have about 3.9 fewer promotion points than their matches due to their lower fitness performance. To translate into a more salient measure, 3.9 points is equivalent to having 0.6 fewer months of experience ($3.9/7.0 = 0.6$). The cumulative effect of job performance is similar to having 1.1 fewer months of experience, and the effect of marksmanship is equivalent to having 0.7 fewer months. The calculation rules use the previous score for fitness tests if a mother has a medical waiver for a given cycle. We assume only one physical fitness test (e.g., PFT) is taken at $r = 8$ and that the second (e.g., CFT) will be taken in the next 6-month cycle (and so will not yet be lower due to postbirth changes). Marines receive ProCons about twice per year. We estimate the effect across $r = \{-4, 2, 8\}$ for someone who was promoted at $r = -10$. To estimate an effect at $r = 12$, we take the effects at $r = \{8, 12\}$ for fitness, $r = \{-6, 1, 6, 12\}$ for job performance (because we drop $r = 0$), and $r = 12$ for rifle scores. Then, the effect at $r = 12$ is equivalent to 0.8 fewer months of experience for fitness, 1.2 fewer months for job performance, and 0.7 fewer months for rifle scores, a total equivalent to 2.7 fewer months of experience.