WHAT DO WORKPLACE WELLNESS PROGRAMS DO? EVIDENCE FROM THE ILLINOIS WORKPLACE WELLNESS STUDY*

DAMON JONES
DAVID MOLITOR
JULIAN REIF

Workplace wellness programs cover over 50 million U.S. workers and are intended to reduce medical spending, increase productivity, and improve well-being. Yet limited evidence exists to support these claims. We designed and implemented a comprehensive workplace wellness program for a large employer and randomly assigned program eligibility and financial incentives at the individual level for nearly 5,000 employees. We find strong patterns of selection: during the year prior to the intervention, program participants had lower medical expenditures and healthier behaviors than nonparticipants. The program persistently increased health screening rates, but we do not find significant causal effects of treatment on total medical expenditures, other health behaviors, employee productivity, or self-reported health status after more than two years. Our 95% confidence intervals rule out 84% of previous estimates on medical spending and absenteeism. *JEL* Codes: I1, M5, J3.

*This research was supported by the National Institute on Aging of the National Institutes of Health under award number R01AG050701; the National Science Foundation under Grant No. 1730546; the Abdul Latif Jameel Poverty Action Lab (J-PAL) North America U.S. Health Care Delivery Initiative; Evidence for Action (E4A), a program of the Robert Wood Johnson Foundation; and the W.E. Upjohn Institute for Employment Research. This study was preregistered with the American Economics Association RCT Registry (AEARCTR-0001368) and was approved by the Institutional Review Boards of the National Bureau of Economic Research, University of Illinois at Urbana-Champaign, and University of Chicago. We thank our coinvestigator Laura Payne for her vital contributions to the study, Lauren Geary for outstanding project management, Michele Guerra for excellent programmatic support, and Illinois Human Resources for invaluable institutional support. We are also thankful for comments from Kate Baicker, Jay Bhattacharya, Tatyana Deryugina, Joseph Doyle, Amy Finkelstein, Colleen Flaherty Manchester, Eliza Forsythe, Drew Hanks, Bob Kaestner, David Meltzer, Michael Richards, Justin Sydnor, Richard Thaler, and numerous seminar participants. We are grateful to Andy de Barros for thoroughly replicating our analysis and to J-PAL for coordinating this replication effort. The findings and conclusions expressed are solely those of the authors and do not represent the views of the National Institutes of Health, any of our funders, or the University of Illinois.

The Quarterly Journal of Economics (2020), 1747–1791. doi:10.1093/qje/qjz023.
Advance Access publication on August 16, 2019.

I. Introduction

Sustained growth in medical spending has prompted policy makers, insurers, and employers to search for ways to reduce health care costs. One widely touted solution is to increase the use of "wellness programs," interventions designed to encourage preventive care and discourage unhealthy behaviors, such as inactivity or smoking. The 2010 Affordable Care Act (ACA) encourages firms to adopt wellness programs by letting them offer participation incentives up to 30% of the total cost of health insurance coverage, and 18 states currently include some form of wellness incentives as a part of their Medicaid program (Saunders et al. 2018). Workplace wellness industry revenue has more than tripled in size to $8 billion since 2010, and wellness programs now cover over 50 million U.S. workers (Mattke, Schnyer, and Van Busum 2012; Kaiser Family Foundation 2016). A meta-analysis by Baicker, Cutler, and Song (2010) finds large medical and absenteeism cost savings, but other studies find only limited benefits (e.g., Gowrisankaran et al. 2013; Baxter et al. 2014). Most of the prior evidence has relied on voluntary firm and employee participation in workplace wellness, limiting the ability to infer causal relationships.

Moreover, the prior literature has generally overlooked important questions regarding selection into wellness programs. If there are strong patterns of selection, the increasing use of large financial incentives now permitted by the ACA may redistribute resources across employees in a manner that runs counter to the intentions of policy makers.[1] For example, wellness incentives may shift costs onto unhealthy or lower-income employees if these groups are less likely to participate. Furthermore, wellness programs may act as a screening device by encouraging healthy employees to join or remain at the firm—perhaps by earning rewards for continuing their healthy lifestyles.

This article investigates two research questions. First, which types of employees participate in wellness programs? While healthy employees may have low participation costs, unhealthy employees may gain the most from participating. Second, what are the causal effects, negative or positive, of workplace wellness

1. Kaiser Family Foundation and Health Research & Educational Trust (2017) estimates that 13% of large firms (at least 200 employees) offer incentives that exceed $500 a year and 4% of large firms offer incentives that exceed $1,000 a year.

programs on medical spending, employee productivity, health behaviors, and well-being? For example, medical spending could decrease if wellness programs improve health or increase if wellness programs and primary care are complements.

To improve our understanding of workplace wellness programs, we designed and implemented the Illinois Workplace Wellness Study, a randomized controlled trial (RCT) conducted at the University of Illinois at Urbana-Champaign (UIUC).[2] We developed a comprehensive workplace wellness program, iThrive, which ran for two years and included three main components: an annual on-site biometric health screening, an annual online health risk assessment (HRA), and weekly wellness activities. We invited 12,459 benefits-eligible university employees to participate in our study and successfully recruited 4,834 participants, 3,300 of whom were assigned to the treatment group and were invited to take paid time off to participate in the wellness program.[3] The remaining 1,534 subjects were assigned to a control group, which was not permitted to participate. Those in the treatment group who successfully completed the entire two-year program earned rewards ranging from $50 to $650, with the amounts randomly assigned and communicated at the start of each program year.

Our analysis combines individual-level data from online surveys, university employment records, health insurance claims, campus gym visit records, and running event records. These data allow us to examine many novel outcomes in addition to the usual ones studied by the prior literature (medical spending and employee absenteeism). From our analysis, we find evidence of significant advantageous selection into our program based on medical spending and health behaviors. At baseline, average annual medical spending among participants was $1,384 less than among nonparticipants. This estimate is statistically ($p = .027$) and economically significant: all else equal, it implies that increasing the share of participating (low-spending) workers employed at

2. Supplemental materials, data sets, and additional publications from this project will be made available on the study website at http://www.nber.org/workplacewellness.

3. UIUC administration provided access to university data and guidance to ensure that our study conformed with university regulations but did not otherwise influence the design of our intervention. Each component of the intervention, including the financial incentives paid to employees, was externally funded.

the university by 4.3 percentage points or more would offset the entire costs of our intervention. Participants were also more likely to have visited campus recreational facilities and to have participated in running events prior to our study. We find evidence of adverse selection when examining productivity: at baseline, participants were more likely to have taken sick leave and were less likely to have worked more than 50 hours a week than were nonparticipants.

Despite strong program participation, we do not find significant effects of our intervention on 40 out of the 42 outcomes we examine in the first year following random assignment. These 40 outcomes include all our measures of medical spending, productivity, health behaviors, and self-reported health. We fail to find significant treatment effects on average medical spending, on different quantiles of the spending distribution, or on any major subcategory of medical utilization (pharmaceutical drugs, office, or hospital). We find no effects on productivity, whether measured using administrative variables (sick leave, salary, promotion), survey variables (hours worked, job satisfaction, job search), or an index that combines all available measures. We also do not find effects on visits to campus gym facilities or on participation in a popular annual community running event, two health behaviors a motivated employee might change within one year. These null effects persist when we estimate longer-run effects of the two-year intervention using outcomes measured up to 30 months after the initial randomization.

Our null estimates are meaningfully precise. For medical spending and absenteeism, two focal outcomes in the prior literature, the 95% confidence intervals of our estimates rule out 84% of the effects reported in 112 prior studies. The 99% confidence interval for the return on investment (ROI) of our intervention rules out the widely cited medical spending and absenteeism ROIs reported in the meta-analysis of Baicker, Cutler, and Song (2010). In addition, our ordinary least squares (OLS) (non-RCT) medical spending estimate, which compares participants with nonparticipants rather than treatment to control, agrees with estimates from prior observational studies. However, the OLS estimate is ruled out by the 99% confidence interval of our instrumental variables (IV) (RCT) estimate. These contrasting results demonstrate the value of using an RCT design in this literature.

Our intervention had two positive treatment effects in the first year, based on responses to follow-up surveys.[4] First, employees in the treatment group were more likely than those in the control group to report ever receiving a health screening. This result indicates that the health screening component of our program did not merely crowd out health screenings that would have otherwise occurred without our intervention. Second, treatment group employees were more likely to report that management prioritizes worker health and safety, although this effect disappears after the first year.

Wellness programs may act as a profitable screening device if they allow firms to preferentially recruit or retain employees with attractive characteristics, such as low health care costs. Prior studies have shown compensation packages can be used in this way (Lazear 2000; Liu et al. 2017), providing additional economic justification for the prevalent and growing use of nonwage employment benefits (Oyer 2008). We find that participation is correlated with preexisting healthy behaviors and low medical spending. However, our estimated retention effects are null after 30 months, which limits the ability of wellness programs to screen employees in our setting.

The results speak to the distributional consequences of workplace wellness. For example, when incentives are linked to pooled expenses such as health insurance premiums, wellness programs may increase insurance premiums for unhealthy low-income workers (Volpp et al. 2011; Horwitz, Kelly, and DiNardo 2013; McIntyre et al. 2017). The results of our selection analysis provide support for these concerns: nonparticipating employees are more likely to be in the bottom quartile of the salary distribution, are less likely to engage in healthy behaviors, and have higher medical expenditures.

We also contribute to the health literature evaluating the causal effects of workplace wellness programs. Most prior studies of wellness programs rely on observational comparisons between participants and nonparticipants (see Pelletier 2011; Chapman 2012, for reviews). Publication bias could skew the set of existing results (Baicker, Cutler, and Song 2010; Abraham and White 2017). To that end, our intervention, empirical specifications, and

4. We address the multiple inference concern that arises when testing many hypotheses by controlling for the family-wise error rate. We discuss our approach in greater detail in Section III.C.

outcome variables were prespecified and publicly archived.[5] Our analyses were also independently replicated by a Jameel Poverty Action Lab (J-PAL) North America researcher.

A number of RCTs have focused on components of workplace wellness, such as wellness activities (Volpp et al. 2008; Charness and Gneezy 2009; Royer, Stehr, and Sydnor 2015; Handel and Kolstad 2017), HRAs (Haisley et al. 2012), or on particular outcomes such as obesity or health status (Meenan et al. 2010; Terry et al. 2011). By contrast, our setting features a comprehensive wellness program that includes a biometric screening, HRA, wellness activities, and financial incentives.

Our study complements the contemporaneous work by Song and Baicker (2019) of a comprehensive wellness program. Similar to us, Song and Baicker (2019) do not find effects on medical spending or employment outcomes after 18 months. Relative to Song and Baicker (2019), our study emphasizes selection into participation, explores in detail the differences between RCT and observational estimates, and includes a longer postperiod (30 months). In contrast to our study, which randomizes at the individual level, Song and Baicker (2019) randomize at the worksite level to capture potential site-level effects, such as spillovers between coworkers. The similarity in results between the two studies—and their divergence from prior work—further underscores the value of RCT evidence within this literature. In addition, our finding that observational estimates are biased toward finding positive health impacts—even after extensive covariate adjustment—reinforces the general concerns about selection bias in observational health studies raised by Oster (2019).

The rest of the article proceeds as follows. Section II provides a background on workplace wellness programs, a description of our experimental design, and a summary of our data sets. Section III outlines our empirical methods, while Section IV presents the results of our first-year analysis. Section V presents results from our longer-run analysis, and Section VI concludes. Finally, all appendix materials can be found in the Online Appendix.

5. Our preanalysis plan is available at http://www.socialscienceregistry.org/trials/1368. We indicate the few instances in which we conduct analyses that were not prespecified. A small number of prespecified analyses have been omitted from the main text for the sake of brevity and because their results are not informative.

II. Experimental Design

II.A. Background

Workplace wellness programs are employer-provided efforts to "enhance awareness, change behavior, and create environments that support good health practices" (Aldana 2001, 297). For the purposes of this study, "wellness programs" encompass three major types of interventions: (i) biometric screenings, which provide clinical measures of health; (ii) HRAs, which assess lifestyle health habits; and (iii) wellness activities, which promote a healthy lifestyle by encouraging specific behaviors (such as smoking cessation, stress management, or fitness). Best practice guides advise employers to let employees take paid time off to participate in wellness programs and to combine wellness program components to maximize their effectiveness (Ryde et al. 2013). In particular, it is recommended that information from a biometric screening and an HRA help determine which wellness activity a person selects (Soler et al. 2010).

Wellness programs vary considerably across employers. Among firms with 200 or more employees, the share offering a biometric screening, HRA, or wellness activities in 2016 was 53%, 59%, and 83%, respectively (Kaiser Family Foundation and Health Research & Educational Trust 2016). These benefits are often coupled with financial incentives for participation, such as cash compensation or discounted health insurance premiums. A 2015 survey estimates an average cost of $693 per employee for these programs (Jaspen 2015), and a recent industry analysis estimates annual revenues of $8 billion (Kaiser Family Foundation 2016).

Several factors may explain the increasing popularity of workplace wellness programs. First, some employers believe that these programs reduce medical spending and increase productivity. For example, Safeway famously attributed its low medical spending to its wellness program (Burd 2009), although this evidence was subsequently disputed (Reynolds 2010). Other work suggests wellness programs may increase productivity (Gubler, Larkin, and Pierce 2017). Second, if employees have a high private value of wellness-related benefits, then labor market competition may drive employers to offer wellness programs to attract and retain workers. Third, the ACA has relaxed constraints on the maximum size of financial incentives offered by employers. Prior to the ACA, health-contingent incentives could not exceed 20% of the cost of employee health coverage. The ACA increased

that general limit to 30% and raised it to 50% for tobacco cessation programs (Cawley 2014). The average premium for a family insurance plan in 2017 was $18,764 (Kaiser Family Foundation and Health Research & Educational Trust 2017), which means that many employers can offer wellness rewards or penalties in excess of $5,000.

Like other large employers, many universities also have workplace wellness programs. Of the nearly 600 universities and liberal arts colleges ranked by *U.S. News & World Report*, more than two-thirds offer an employee wellness program.[6] Prior to our intervention, UIUC's campus wellness services were run by the University of Illinois Wellness Center, which has one staff member. The Wellness Center only coordinates smoking cessation resources for employees and provides a limited number of wellness activities, many of which are not free. Importantly for our study, the campus did not offer any health screenings or HRAs and did not provide monetary incentives to employees in exchange for participating in wellness activities. Therefore, our intervention effectively represents the introduction of all major components of a wellness program at this worksite.

II.B. The Illinois Workplace Wellness Study and iThrive

The Illinois Workplace Wellness Study is a large-scale RCT designed to investigate the effects of workplace wellness programs on employee medical spending, productivity, and well-being. As part of the study, we worked with the director of Campus Wellbeing Services to design and to introduce a comprehensive wellness program named iThrive at UIUC. Our goal was to create a representative program that includes all the key components recommended by wellness experts: a biometric screening, an HRA, a variety of wellness activities, monetary incentives, and paid time off. We summarize the program here and provide full details in Online Appendix D.

Figure I illustrates the experimental design of the first year of our study. In July 2016 we invited 12,459 benefits-eligible university employees to enroll in our study by completing a 15-minute online survey designed to measure baseline health and wellness

6. Source: authors' tabulation of data collected from universities and colleges via website search and phone inquiry.

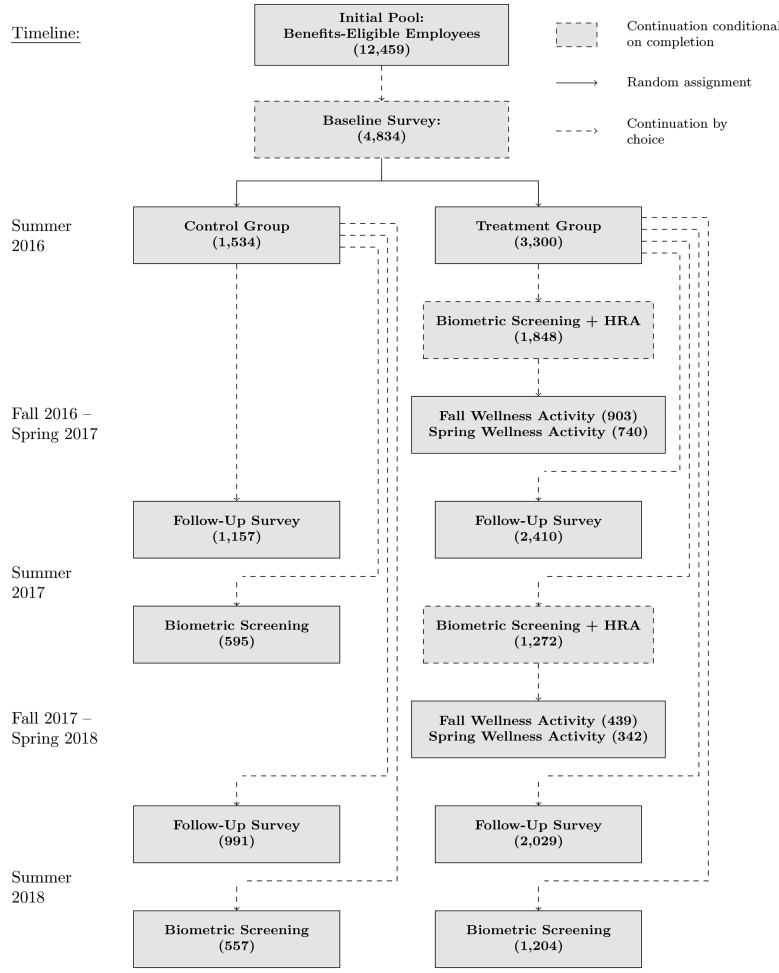


FIGURE I

Experimental Design of the Illinois Workplace Wellness Study

(dependents were not eligible to participate).[7] The invitations were sent by postcard and email; employees were offered a $30 Amazon.com gift card to complete the survey as well as a chance "to participate in a second part of the research study." Over the

7. Participation required providing informed consent and completing the online baseline survey.

course of three weeks, 4,834 employees completed the survey. Study participants, whom we define as anybody completing the survey, were then randomly assigned to either the control group ($N = 1{,}534$) or the treatment group ($N = 3{,}300$). Members of the control group were notified that they may be contacted for follow-up surveys in the future, and further contact with this group was thereafter minimized. Members of the treatment group were offered the opportunity to participate in iThrive.

The first step of iThrive included a biometric health screening and an online HRA. For five weeks in August and September 2016, participants could schedule a screening at one of many locations on campus. A few days after the screening, they received an email invitation to complete the online HRA designed to assess their lifestyle habits. Upon completing it, participants were given a score card incorporating the results of their biometric screening and providing recommended areas of improvement. Only participants who completed both the screening and HRA were eligible to participate in the second step of the program.

The second step of iThrive consisted of wellness activities. Eligible participants were offered the opportunity to participate in one of several activities in the fall semester and another activity in the spring semester. Eligibility to participate in the spring activities was not contingent on enrollment or completion of fall activities. In the fall, activities included in-person classes on chronic disease management, weight management, tai chi, physical fitness, financial wellness, and healthy workplace habits; a tobacco quitline; and an online, self-paced wellness challenge. A similar set of activities was offered in the spring. Classes ranged from 6 to 12 weeks in length, and "completion" of a class was generally defined as attending at least three-fourths of the sessions. Participants were given two weeks to enroll in wellness activities and were encouraged to incorporate their HRA feedback when choosing a class.

Study participants were offered monetary rewards for completing each step of the iThrive program, and these rewards varied depending on the treatment group to which they were assigned. Individuals in treatment groups A, B, and C were offered a screening incentive of \$0, \$100, or \$200, respectively, for completing the biometric screening and the HRA in the first year. Treatment groups were further split based on an activity incentive of either \$25 or \$75 for each wellness activity completed (up

to one per semester). Thus, there were six treatment groups in total: A25, A75, B25, B75, C25, and C75 (see Online Appendix Figure D.1).

The total reward for completing all iThrive components—the screening, HRA, and a wellness activity during both semesters—ranged from $50 to $350 in the first year, depending on the treatment group. These amounts are in line with typical wellness programs (Mattke, Schnyer, and Van Busum 2012). The probability of assignment to each group was equal across participants, and randomization was stratified by employee class (faculty, staff, or civil service), sex, age, quartile of annual salary, and race (see Online Appendix D.1.2 for additional randomization details). We privately informed participants about their screening and wellness activity rewards at the start of the intervention (August 2016) and did not disclose information about rewards offered to others.

To help guide participants through iThrive, we developed a secure website with information about the program. At the onset of iThrive in August 2016, the website instructed participants to schedule a biometric screening and then take the online HRA. Beginning in October 2016, and then again in January 2017, the website provided a menu of wellness activities and online registration forms for those activities as well as information on their current progress and rewards earned to date, answers to frequently asked questions, and contact information for support.

We implemented a second year of our intervention beginning in August 2017. As in the first year, treatment group participants were offered a biometric screening, an HRA, and various wellness activities (see Online Appendix Figure D.2 for more details). Our study concluded with a third and final health screening in August 2018. For comparison purposes, we invited both the treatment and control groups to complete all follow-up surveys and screenings in 2017 and 2018. We discuss the second-year intervention in more detail in Section V.

II.C. Data

For our analysis, we link together several survey and administrative data sets at the individual level. Each data source is summarized in this section and detailed in Online Appendix Section D.2. Online Appendix Table A.14 defines each variable used in the analysis and notes which outcomes were not prespecified.

1. University Administrative Data. We obtained university administrative data on 12,459 employees who, as of June 2016, were (i) working at the Urbana-Champaign campus at the University of Illinois and (ii) were eligible for part- or full-time employee benefits from the Illinois Department of Central Management Services. The initial denominator file includes employee name, university identification number, contact information (email and home mailing address), date of birth, sex, race, job title, salary, and employee class (faculty, academic staff, or civil service). We used email and home mailing addresses to invite employees to participate in our study, and we used sex, race, date of birth, salary, and employee class to generate the strata for random sampling.

A second file includes employment history information as of July 31, 2017. This file provides three employment and productivity outcomes measured over the first 12 months of our study: job termination date (for any reason, including firings or resignations), job title change (since June 2016), and salary raises. The average salary raise in our main sample was 5.9% after one year. For those with a job title change in the first year, the average raise was 14.5%, and a small number (<5%) of employees with job title changes did not receive an accompanying salary raise. We also define an additional variable, "job promotion," which is an indicator for receiving both a title change and a salary raise, thus omitting title changes that are potentially lateral moves or demotions.[8] We obtained an updated version of this employment history file on January 31, 2019, for the longer-run analysis presented in Section V.

A third file provides data on sick leave. The number of sick days taken is available at the monthly level for civil service employees; for academic faculty and staff, the number of sick days taken is available biannually, on August 15 and May 15. We first calculate the total number of sick days taken during our preperiod (August 2015–July 2016) and postperiod (August 2016–July 2017) for each employee. We then normalize by the number of days employed to make this measure comparable across employees. All specifications that include sick days taken as an outcome variable are weighted by the number of days employed. Our longer-run analysis, presented in Section V, uses a newer version of this file that includes a postperiod covering August 2016–January 2019.

8. We did not prespecify the job promotion or job title change outcomes in our preanalysis plan.

A fourth file contains data on exact attendance dates for the university's gym and recreational facilities. Entering one of these facilities requires swiping an ID card, which creates a database record linked to the person's university ID. We calculate the total number of visits per year for the preperiod and the postperiod. As with the sick leave data, our longer-run analysis uses a version of this file that includes the postperiod.

2. Online Survey Data. As described in Section II.B, all study participants took a 15-minute online baseline survey in July 2016 as a condition of enrollment in the study. The survey covered topics including health status, health care use, job satisfaction, and productivity.

Our survey software recorded that, out of the 12,459 employees invited to take the survey, 7,468 clicked on the link to the survey, 4,918 began the survey, and 4,834 completed the survey. Although participants were allowed to skip questions, response rates for the survey were very high: 4,822 out of 4,834 participants (99.7%) answered every question used in our analysis. To measure the reliability of the survey responses, we included a question about age at the end of the survey and compared participants' self-reported ages with the ages listed in the university's administrative data. Of the 4,830 participants who reported an age, only 24 (<0.5%) reported a value that differed from the university's administrative records by more than one year.

All study participants were invited via postcard and email to take a one-year follow-up survey online in July 2017.[9] In addition to the questions asked on the online baseline survey, the follow-up survey included additional questions on productivity, presenteeism, and job satisfaction. A total of 3,567 participants (74%) successfully completed the 2017 follow-up survey. The completion rates for the control and treatment groups were 75.4% and 73.1%, respectively. The difference in completion rates is small but marginally significant ($p = .079$).

Finally, we invited all study participants to take a two-year follow-up survey in July 2018. In total, 3,020 participants (62.5%) completed the survey. The completion rates for the control and treatment groups were 64.6% and 61.5%, respectively. The completion rate difference remains small but becomes more

9. Invitations to the follow-up survey were sent regardless of current employment status with the university.

statistically significant ($p = .036$). Full texts of our surveys are available in our supplementary materials.[10]

3. Health Insurance Claims Data. We obtained health insurance claims data for January 1, 2015, through July 31, 2017, for the 67% of employees who subscribe to the university's most popular insurance plan. We use the total payment due to the provider to calculate average total monthly spending. We also use the place of service code on the claim to break total spending into four major subcategories: pharmaceutical, office, hospital, and other.[11] Our spending measures include all payments from the insurer to providers as well as any deductibles or co-pays paid by individuals.

Employees choose their health plan annually during May, and plan changes become effective July 1. Participants were informed of their treatment assignment on August 9, 2016. We therefore define baseline medical spending to include all allowed amounts with dates of service corresponding to the 13-month time period of July 1, 2015, through July 31, 2016. We define spending in the postperiod to correspond to the 12-month time period of August 1, 2016, through July 31, 2017. For the longer-run analysis presented in Section V, we obtained an updated version of the claims file that allowed us to define a postperiod corresponding to the 30-month period August 1, 2016, through January 31, 2019.

In our health claims sample, 11% of employees are not continuously enrolled throughout the 13-month preperiod, and 9% are not continuously enrolled throughout the 12-month postperiod, primarily due to job turnover. Because average monthly spending is measured with less noise for employees with more months of claims, we weight regressions by the number of covered months whenever the outcome variable is average spending.

4. Illinois Marathon/10K/5K Data. The Illinois Marathon is a running event held annually in Champaign. The individual races are a marathon, a half marathon, a 5K, and a 10K.

10. Interactive versions of the study surveys are available at http://www.nber.org/workplacewellness.

11. Pharmaceutical and office-based spending each have their own place of service codes. Hospital spending is summed across the following four codes: "Off-Campus Outpatient Hospital," "Inpatient Hospital," "On-Campus Outpatient Hospital," and "Emergency Room Hospital." All remaining codes are assigned to "other" spending, which serves as the omitted category in our analysis.

When registering for a race, a participant must provide their name, age, sex, and hometown. That information, along with the results of the race, are published online after the races have ended. We downloaded those data for the 2014–2018 races and matched it to individuals in our data set using name, age, sex, and hometown.

5. Employee Productivity Index. To help measure productivity, we construct an index equal to the first principal component of all survey and administrative measures of employee productivity. Online Appendix Table A.8 shows that this index depends negatively on sick leave and likelihood of job search and positively on salary raises, job satisfaction, and job promotion.

II.D. Baseline Summary Statistics and Balance Tests

Table I provides baseline summary statistics for the employees in our sample. Columns (2) and (3) report means for those who were assigned to the control and treatment groups, respectively. Column (1) reports means for employees not enrolled in our study, as available. The variables are grouped into four panels, based on the source and type of data. Panel A presents means of the university administrative data variables used in our stratified randomization, Panel B presents means of variables from our 2016 online baseline survey, Panel C presents means of medical spending variables from our health insurance claims data for the July 2015–July 2016 time period, and Panel D presents baseline means of administrative data variables used to measure health behaviors and employee productivity.

Our experimental framework relies on the random assignment of study participants to treatment. To evaluate the validity of this assumption, we test whether the control and treatment means are equal and whether the variables listed within each panel jointly predict treatment assignment.[12] By construction, we find no evidence of differences in means among the variables used for stratification (Panel A): all p-values in column (4) are greater than .7. Among all other variables listed in Panels B, C, and D, we find statistically significant differences at a 10% or lower level in 2 out of 34 cases, which is approximately what one would expect from random chance. Our joint balance tests fail to reject the null

12. Online Appendix Tables A.1a and A.1b report balance tests across subtreatment arms.

TABLE I

MEANS OF STUDY VARIABLES AT BASELINE

	Not in study (1)	Enrolled in study: Control (2)	Enrolled in study: Treatment (3)	*p*-value (4)	Sample size (5)
Panel A: Stratification variables					
Male	0.536	0.426	0.428	.902	12,459
Age 50+	0.430	0.323	0.327	.818	12,459
Age 37–49	0.362	0.340	0.332	.591	12,459
White	0.774	0.841	0.836	.648	12,459
Salary Q1 (bottom quartile)	0.234	0.244	0.242	.881	12,459
Salary Q2	0.189	0.255	0.259	.773	12,459
Salary Q3	0.197	0.249	0.250	.924	12,459
Faculty	0.298	0.196	0.201	.721	12,459
Academic staff	0.324	0.443	0.437	.712	12,459
Panel B: 2016 survey variables					
Ever screened		0.885	0.892	.503	4,834
Physically active		0.359	0.382	.134	4,834
Trying to be active		0.822	0.809	.278	4,834
Current smoker (cigarettes)		0.072	0.065	.428	4,833
Current smoker (other)		0.085	0.085	.960	4,833
Former smoker		0.198	0.196	.870	4,833
Drinker		0.657	0.645	.423	4,830
Heavy drinker		0.050	0.049	.955	4,829
Chronic condition		0.729	0.726	.824	4,834
Excellent or v. good health		0.586	0.602	.281	4,834
Not poor health		0.989	0.989	.882	4,834
Physical problems		0.392	0.388	.793	4,834
Lots of energy		0.310	0.330	.171	4,834
Bad emotional health		0.308	0.288	.162	4,834
Overweight		0.545	0.533	.438	4,834
High BP/cholesterol/glucose		0.308	0.295	.354	4,834
Sedentary		0.545	0.542	.846	4,833
Pharmaceutical drug utilization		0.723	0.706	.205	4,830
Physician/ER utilization		0.772	0.748	.077	4,833
Hospital utilization		0.038	0.027	.054	4,833
Any sick days in past year		0.618	0.600	.232	4,828
Worked 50+ hours/week		0.187	0.173	.234	4,831
Very satisfied with job		0.396	0.408	.415	4,832
Very or somewhat satisfied with job		0.836	0.845	.419	4,832
Management priority on health/safety		0.771	0.782	.401	4,831
Sample size	7,625	1,534	3,300		
Joint balance test for Panel A (*p*-value)				1.000	4,834
Joint balance test for Panel B (*p*-value)				.821	4,817

TABLE I
(CONTINUED)

	Not in study (1)	Enrolled in Study: Control (2)	Enrolled in Study: Treatment (3)	*p*-value (4)	Sample size (5)
Panel C: Health claims variables (2015–2016)					
Total spending (dollars/month)	579	506	465	.317	8,096
Office spending	54	67	58	.498	8,096
Hospital spending	345	283	259	.369	8,096
Drug spending	105	103	101	.911	8,096
Nonzero medical spending	0.888	0.899	0.885	.253	8,096
Panel D: Health behavior and productivity variables					
Sick leave (days/year)	5.89	6.05	6.13	.707	12,459
Annual salary (dollars)	73,927	61,528	61,736	.840	12,221
IL Marathon/10K/5K (2014–2016)	0.072	0.107	0.118	.257	12,459
Campus gym visits (days/year)	6.14	7.36	6.78	.483	12,459
Sample size	7,625	1,534	3,300		
Joint balance test for Panel C (*p*-value)				.764	3,223
Joint balance test for Panel D (*p*-value)				.752	4,770

Notes. Columns (1)–(3) report unweighted means for different, nonoverlapping subsets of university employees. Column (4) reports the *p*-value from a joint test of equality of the two coefficients reported in columns (2)–(3). The joint balance test rows report the *p*-value from jointly testing whether the variables in a particular panel predict enrollment into treatment.

hypothesis that the variables in Panel B ($p = .821$), Panel C ($p = .764$), or Panel D ($p = .752$) are not predictive of assignment to treatment.

A unique feature of our study is our ability to characterize the employees who declined to participate in the experiment. We investigate the extent of this selection into our study by comparing means for study participants, reported in Table I, columns (2)–(3), to the means for nonparticipating employees who did not complete our online baseline survey, reported in column (1). Study participants are younger, are more likely to be female, are more likely to be white, have lower incomes on average, are more likely to be administrative staff, and are less likely to be faculty. They also have lower baseline medical spending, are more likely to have participated in one of the Illinois Marathon/10K/5K running events, and have a higher rate of monthly gym visits. These selection effects mirror the ones we report in Section IV.B, suggesting that the factors governing the decision to participate in a wellness program are similar to the ones driving the decision to participate in our study.

III. Empirical Methods

III.A. Selection

First, we characterize the types of employees who are most likely to complete the various stages of our wellness program in the first year. We estimate the following OLS regression using observations from the treatment group:

$$X_i = \alpha + \theta P_i + \varepsilon_i. \tag{1}$$

The left-hand-side variable, X_i, is a predetermined covariate. The regressor, P_i, is an indicator for one of the following three participation outcomes: completing a screening and HRA, completing a fall wellness activity, or completing a spring wellness activity. The coefficient θ represents the correlation between participation and the baseline characteristic, X_i; it should not be interpreted causally.

III.B. Causal Effects

Next we estimate the effect of our wellness intervention on a number of outcomes, including medical spending from health claims data, employment and productivity variables measured in administrative and survey data, health behaviors measured in administrative data, and self-reported health status and behaviors. We compare outcomes in the treatment group to those in the control group using the following specification:

$$Y_i = \alpha + \gamma T_i + \Gamma X_i + \varepsilon_i. \tag{2}$$

Here, T_i is an indicator for membership in the treatment group, and Y_i is an outcome of interest. We estimate equation (2) with and without the inclusion of controls, X_i. In one control specification, X_i includes baseline strata fixed effects. One could also include a much broader set of controls, but doing so comes at the cost of reduced degrees of freedom. Thus, our second control specification implements the Lasso double-selection method of Belloni, Chernozhukov, and Hansen (2014), as outlined by Urminsky, Hansen, and Chernozhukov (2016), which selects controls that predict either the dependent variable or the focal independent variable.[13]

13. No control variable will be predictive of a randomly assigned variable, in expectation. Thus, when implementing the double-selection method with randomly

The set of potential controls includes baseline values of the outcome variable, strata variables, the baseline survey variables reported in Table I, and all pairwise interactions. We then estimate a regression that includes only the controls selected by double-Lasso. In our tables, we follow convention and refer to this third control strategy as "post-Lasso." As before, our main identifying assumption requires treatment to be uncorrelated with unobserved determinants of the outcome. The key parameter of interest, γ, is the intent-to-treat (ITT) effect of our intervention on the outcome Y_i.

III.C. Inference

We report conventional robust standard errors in all tables. We do not cluster standard errors because randomization was performed at the individual level (Abadie et al. 2017). Because we estimate equations (1) and (2) for many different outcome variables, the probability that we incorrectly reject at least one null hypothesis is greater than the significance level used for each individual hypothesis test. When appropriate, we address this multiple inference concern by controlling for the family-wise error rate (i.e., the probability of incorrectly rejecting one or more null hypotheses belonging to a family of hypotheses).

To control for the family-wise error rate, we first define eight mutually exclusive families of hypotheses that encompass all of our outcome variables. Each family contains all variables belonging to one of our four outcome domains (strata variables, medical spending, employment/productivity, or health) and one of our two types of data (administrative or survey).[14] When testing multiple hypotheses using equations (1) and (2), we calculate family-wise

assigned treatment status as the focal independent variable, we only select controls that are predictive of the dependent variable. When implementing Lasso, we use the penalty parameter that minimizes the 10-fold cross-validated mean squared error.

14. One could assign all variables to a single family of hypotheses. This is unappealing, however, because it assigns equal importance to all outcomes when in fact some outcomes (e.g., total medical spending) are of much greater interest than others. Instead, our approach groups together variables that measure related outcomes and that originate from similar data sources. Because it is based on both survey and administrative data, we assign the productivity index variable to its own (ninth) family.

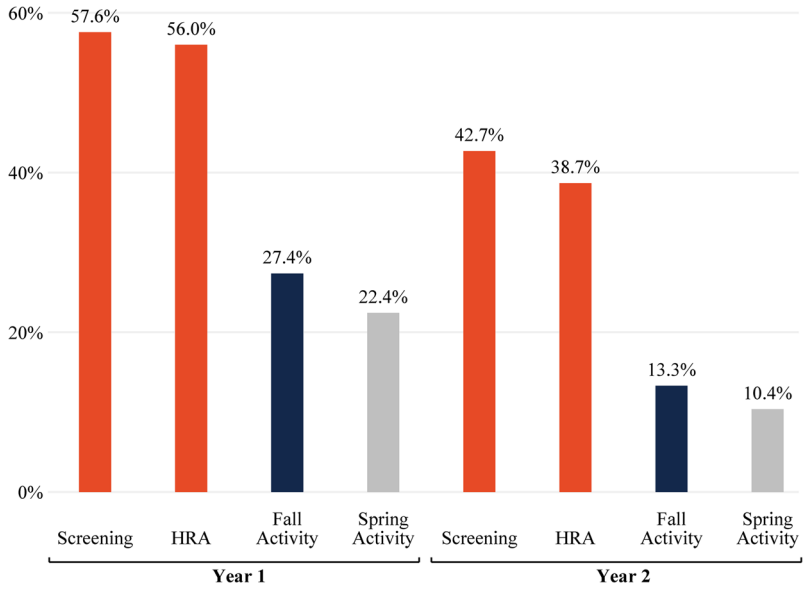


FIGURE II

Employee Participation Rates in the iThrive Workplace Wellness Program

Participation rates are measured as a fraction of the treatment group (N = 3,300).

adjusted p-values based on 10,000 bootstraps of the free step-down procedure of Westfall and Young (1993).[15]

IV. First-Year Results

IV.A. Participation

Figure II reports that 56.0% of participants in the treatment group completed both the health screening and online HRA in the first year. These participants earned their assigned rewards and were allowed to participate in wellness activities; the remaining 44% of the treatment group was not allowed to sign up for these first-year activities. In the fall, 27.4% of the treatment group completed enough of the activity to earn their assigned activity reward. Completion rates were slightly lower (22.4%) for the spring

15. We have made our generalized Stata code module publicly available for other interested researchers to use. It can be installed by typing "`ssc install wyoung, replace`" at the Stata prompt. We provide additional documentation of this multiple testing adjustment in Online Appendix C.

wellness activities. By way of comparison, a survey of employers with workplace wellness programs found that less than 50% of their eligible employees complete health screenings and that most firms have wellness activity participation rates of less than 20% (Mattke et al. 2013). In the second year, participation rates follow a similar qualitative pattern, although the level of participation is shifted down for all activities. This reduction reflects job turnover and may also be due, at least in part, to the smaller size of the rewards offered in the second year.

Except for the second-year screening—which was also offered to the control group—these participation rates quantify the "first-stage" effect of treatment on participation. This is formalized in Online Appendix Table A.2, which reports the first-stage estimates by regressing the completion of the eight steps in Figure II on an indicator for treatment group membership. In our IV specifications, we use completion of the first-year HRA as the relevant participation outcome in the first stage.

IV.B. Selection

1. Average Selection. Next we characterize the types of workers most likely to participate in our wellness program. We report selected results in Table II and present results for the full set of prespecified outcomes in Online Appendix Tables A.3a–A.3d. We test for selection at three different sequential points in the first year of the study: completing the health screening and HRA, completing a fall wellness activity, and completing a spring wellness activity. Column (1) reports the mean of the selection variable of interest for employees assigned to the treatment group, and columns (3)–(5) report the difference in means between those employees who successfully completed the participation outcome of interest and those who did not. We also report family-wise p-values in brackets that account for the number of selection variables in each "family."[16]

Table II, column (3), first row reports that employees who completed the screening and HRA spent, on average, $115.3 per month less on health care in the 13 months prior to our study than

16. The eight families of outcome variables are defined in Section III.C. The family-wise p-values reported in Table II account for all the variables in the family, including ones not reported in the main text. An expanded version of Table II that reports estimates for all prespecified outcomes is provided in Online Appendix Tables A.3a–A.3d.

TABLE II

SELECTION ON MEDICAL SPENDING, PRODUCTIVITY, AND HEALTH BEHAVIORS

Selection variable	Mean (1)	*N* (2)	Completed screening and HRA (3)	Completed fall activity (4)	Completed spring activity (5)
Panel A: Baseline medical spending					
Total spending (dollars/month) [admin]	479	2,188	−115.3**	−60.6	−62.5
			(52.2)	(43.6)	(44.3)
			[.082]	[.405]	[.273]
Nonzero medical spending [admin]	0.885	2,188	0.050***	0.049***	0.046***
			(0.014)	(0.014)	(0.014)
			[.008]	[.005]	[.020]
Panel B: Baseline productivity					
Productivity index [survey/admin]	0.008	3,251	−0.077	−0.099**	−0.104**
			(0.047)	(0.050)	(0.052)
			[.096]	[.046]	[.044]
Sick leave (days/year) [admin]	6.274	3,296	0.473*	0.705**	0.617**
			(0.267)	(0.290)	(0.312)
			[.144]	[.015]	[.048]
Worked 50+ hours/week [survey]	0.173	3,297	−0.058***	−0.065***	−0.064***
			(0.013)	(0.014)	(0.014)
			[.000]	[.000]	[.000]
Annual salary (dollars) [admin]	61,736	3,257	−782.7	−3,363.9***	−3,429.1***
			(1,248.3)	(1,191.6)	(1,251.8)
			[.519]	[.009]	[.012]
Salary Q1 (bottom quartile) [admin]	0.242	3,300	−0.069***	−0.022	−0.036**
			(0.015)	(0.016)	(0.017)
			[.000]	[.398]	[.121]
Panel C: Baseline health behaviors					
IL Marathon/10K/5K (2014–2016) [admin]	0.118	3,300	0.089***	0.111***	0.090***
			(0.011)	(0.014)	(0.016)
			[.000]	[.000]	[.000]
Campus gym visits (days/year) [admin]	6.780	3,300	2.178**	1.006	1.629
			(0.885)	(1.024)	(1.132)
			[.013]	[.328]	[.153]

Notes. Column (1) reports the mean among subjects assigned to treatment. Columns (3)–(5) report the difference in means between those who completed the participation outcome and those who did not. Robust standard errors are reported in parentheses. */**/*** indicates significance at the 10%/5%/1% level using conventional inference (i.e., not adjusting for multiple outcomes). Family-wise *p*-values, reported in brackets, adjust for the number of outcome (selection) variables in each family and are estimated using 10,000 bootstraps.

employees who did not participate. This pattern of advantageous selection is strongly significant using conventional inference ($p = .027$) and remains marginally significant after adjusting for the five outcomes in this family (family-wise $p = .082$). The magnitude is also economically significant, representing 24% of the $479 in average monthly spending (column (1)). Columns (4) and (5) present further evidence of advantageous selection into the fall and spring wellness activities, although in these cases

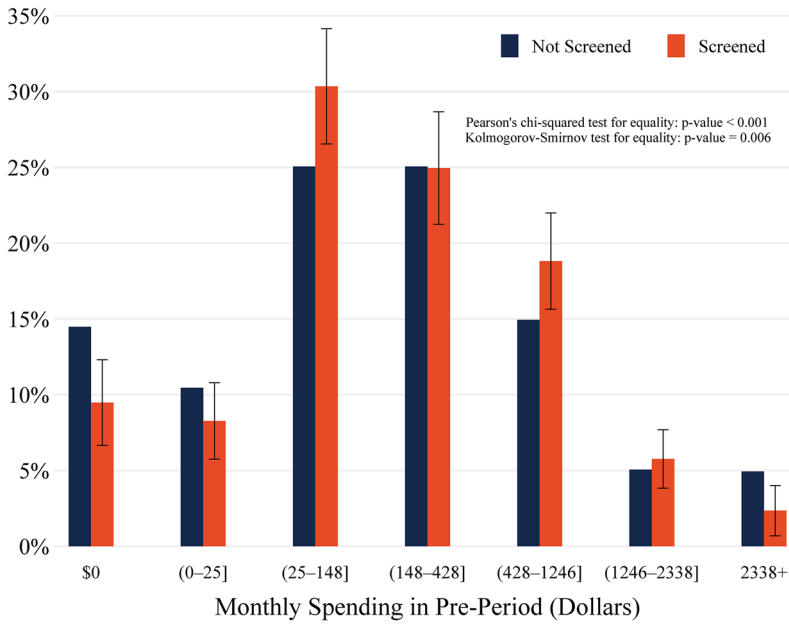


FIGURE III

Preintervention Medical Spending among Treatment Group, by Participation Status

Data are from claims covering the period July 2015–July 2016 ($N = 2{,}188$). The first two bins ($0 and (0–25]) include 25% of those not screened. The remaining five bins were defined to include 25%, 25%, 15%, 5%, and 5% of those not screened, respectively. The null hypothesis of the Pearson's chi-squared and the nonparametric Kolmogorov-Smirnov tests is that the two samples are drawn from the same distribution.

the magnitude of selection falls by half and becomes statistically insignificant.

In contrast, the second row of Table II reports that employees participating in our wellness program were more likely to have nonzero medical spending at baseline than nonparticipants, by about 5 percentage points (family-wise $p \leqslant .02$), for all three participation outcomes. When combined with our results from the first row on average spending, this suggests that our wellness program is more attractive to employees with moderate spending than to those in either tail of the spending distribution.

We investigate these results further in Figure III, which displays the empirical distributions of prior spending for those employees who participated in screening and for those who did not.

Pearson's chi-squared test and the nonparametric Kolmogorov-Smirnov test both strongly reject the null hypothesis that these two samples were drawn from the same distribution (Chi-squared $p < .001$; Kolmogorov-Smirnov $p = .006$).[17] Figure III reveals a "tail-trimming" effect: participating (screened) employees are less likely to be high spenders (>$2,338 a month), but they are also less likely to be low spenders ($0 a month). Because medical spending is right-skewed, the overall effect on the mean among participants is negative, which explains the advantageous selection effect reported in the first row of Table II.

Table II, Panel B reveals negative selection on our productivity index, a summary measure of productivity. This result is driven in part by positive selection on prior sick leave taken and negative selection on working over 50 hours a week and on salary. The average annual salary of participants is lower than that of nonparticipants, significantly so for the fall and spring wellness activities (family-wise $p \leqslant .012$). This initially suggests that participants are disproportionately lower income; yet the share of screening participants in the first (bottom) quartile of income is actually 6.9 percentage points lower than the share among nonparticipants (family-wise $p < .001$). Columns (4) and (5) report negative, albeit smaller, selection effects for the fall and spring wellness activities. We again delve deeper by comparing the entire empirical distributions of income for participants and nonparticipants in Figure IV. We can reject that these two samples came from the same distribution ($p \leqslant .002$). As in Figure III, we again find a tail-trimming effect: participating employees are less likely to come from either tail of the income distribution.

Last, we test for differences in baseline health behaviors as measured by our administrative data variables. The first row of Table II, Panel C reports that the share of screening participants who had previously participated in one of the Illinois Marathon/5K/10K running events is 8.9 percentage points larger than the share among nonparticipants (family-wise $p < .001$), a sizable difference that represents over 75% of the mean participation rate of 11.8% (column (1)). This selection effect is even larger for the fall and spring wellness activities. The second row of Panel C reports that participants also visited the campus gym facilities more frequently, although these selection effects are only statistically significant for screening and HRA completion (family-wise $p = .013$).

17. These tests were not specified in our preanalysis plan.

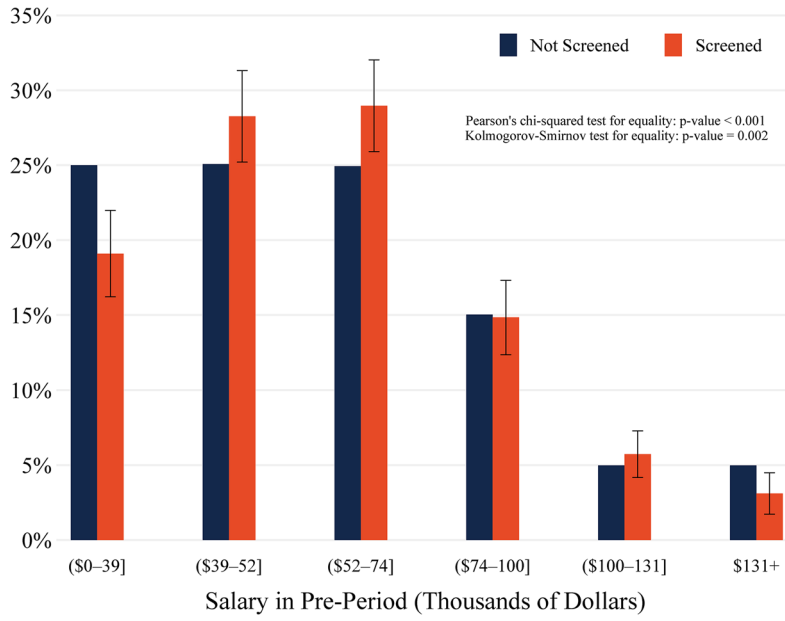


FIGURE IV

Preintervention Salary among Treatment Group, by Participation Status

Salary was measured on June 1, 2016 (N = 3,257). The six bins were defined to include 25%, 25%, 25%, 15%, 5%, and 5% of employees not screened, respectively. The null hypothesis of the Pearson's chi-squared and the nonparametric Kolmogorov-Smirnov tests is that the two samples are drawn from the same distribution.

Prior studies have raised concerns that the benefits of wellness programs accrue primarily to higher-income employees with lower health risks (Horwitz, Kelly, and DiNardo 2013). Our results are broadly consistent with these concerns: participating employees are less likely to have very high medical spending, less likely to be in the bottom quartile of income, and more likely to engage in healthy physical activities. At the same time, participating employees are also less likely to have very low medical spending or have very high incomes, which suggests a more nuanced story. In addition, we find that less productive employees are more likely to participate, particularly in the wellness activity portion of the program, suggesting it may be less costly for these employees to devote time to the program.

2. Health Care Cost Savings via Selection. The selection patterns we have uncovered may provide, by themselves, a potential motive for firms to offer wellness programs. We have shown that wellness participants have lower medical spending on average than nonparticipants. If wellness programs differentially increase the recruitment or retention of these types of employees, then the accompanying reduction in health care costs will save firms money.[18]

A simple back-of-the-envelope calculation demonstrates this possibility. In our setting, 39% $\left(=\frac{4,834}{12,459}\right)$ of eligible employees enrolled into our study, and 56% of the treatment group completed a screening and health assessment (Figure II). Participating employees spent, on average, \$138.2 a month less than nonparticipants in the postperiod (Table IV, column (4)), which translates into an annual spending difference of \$1,658. When combined with average program costs of \$271 per participant, this implies that the employer would need to increase the share of employees who are similar to wellness participants by 4.3 $\left(\text{e.g., } \frac{0.39\times0.56\times271}{1,658-271}\right)$ percentage points for the resulting reduction in medical spending to offset the entire cost of the wellness program.

To be clear, this calculation does not assume or imply that adoption of workplace wellness programs is socially beneficial. But it does provide a profit-maximizing rationale for firms to adopt wellness programs, even in the absence of any direct effects on health, productivity, or medical spending. Section V, however, shows we do not find any effects on retention after 30 months, so if this effect exists in our setting, then it needs to operate through a recruitment channel, which we cannot estimate using our study design.

IV.C. Causal Effects

1. Intent-to-Treat. We estimate the causal, ITT effect of our intervention on three domains of outcomes: medical spending, employment and productivity, and health behaviors. Table III reports estimates of equation (2) for selected outcomes. An expanded version of this table reporting results for all 42 administrative and

18. Wellness participants differ from nonparticipants along other dimensions as well (e.g., health behaviors). Because it is difficult in many cases to sign (let alone quantify) a firm's preferences over these other dimensions, we focus our cost-savings discussion on the medical spending consequences.

TABLE III

TREATMENT EFFECTS (INTENTION-TO-TREAT)

Outcome variable	First year (12 months)			Longer-run (24–30 months)		
	Mean (1)	No controls (2)	Post-Lasso (3)	Mean (4)	No controls (5)	Post-Lasso (6)
Panel A: Medical spending						
Total spending (dollars/month) [admin]	576.2	10.8	34.9	650.5	−74.7	−39.7
		(48.5)	(36.9)		(58.5)	(47.9)
		[.937]	[.859]		[.618]	[.754]
	$N = 3{,}239$	$N = 3{,}239$	$N = 3{,}152$	$N = 3{,}307$	$N = 3{,}307$	$N = 3{,}155$
Drug spending [admin]	132.0	−8.5	−6.1	148.8	−25.2	−22.2
		(26.5)	(12.0)		(27.7)	(16.4)
		[.937]	[.947]		[.836]	[.589]
	$N = 3{,}239$	$N = 3{,}239$	$N = 3{,}152$	$N = 3{,}307$	$N = 3{,}307$	$N = 3{,}155$
Office spending [admin]	69.5	−6.1	−2.0	74.2	−6.6	−4.8
		(10.0)	(4.4)		(8.6)	(5.4)
		[.937]	[.947]		[.836]	[.754]
	$N = 3{,}239$	$N = 3{,}239$	$N = 3{,}152$	$N = 3{,}307$	$N = 3{,}307$	$N = 3{,}155$
Hospital spending [admin]	313.0	22.2	24.6	353.5	−31.7	−20.3
		(30.9)	(28.1)		(35.6)	(31.9)
		[.937]	[.868]		[.836]	[.754]
	$N = 3{,}239$	$N = 3{,}239$	$N = 3{,}152$	$N = 3{,}307$	$N = 3{,}307$	$N = 3{,}155$
Nonzero medical spending [admin]	0.902	−0.008	0.002	0.950	0.005	0.007
		(0.011)	(0.010)		(0.008)	(0.007)
		[.937]	[.947]		[.836]	[.754]
	$N = 3{,}239$	$N = 3{,}239$	$N = 3{,}152$	$N = 3{,}307$	$N = 3{,}307$	$N = 3{,}155$

TABLE III
(CONTINUED)

Outcome variable	First year (12 months)			Longer-run (24–30 months)		
	Mean (1)	No controls (2)	Post-Lasso (3)	Mean (4)	No controls (5)	Post-Lasso (6)
Panel B: Employment and Productivity						
Job promotion [admin]	0.176	−0.003	−0.004	0.360	0.006	0.006
		(0.013)	(0.012)		(0.017)	(0.016)
		[.952]	[.922]		[.996]	[.996]
	$N = 4,146$	$N = 4,146$	$N = 4,130$	$N = 3,635$	$N = 3,635$	$N = 3,619$
Job terminated [admin]	0.113	−0.013	−0.012	0.204	0.002	0.002
		(0.010)	(0.009)		(0.012)	(0.012)
		[.630]	[.571]		[1.000]	[.998]
	$N = 4,834$	$N = 4,834$	$N = 4,753$	$N = 4,834$	$N = 4,834$	$N = 4,753$
Sick leave (days/year) [admin]	6.341	0.186	0.138	6.066	0.013	0.018
		(0.230)	(0.200)		(0.204)	(0.169)
		[.816]	[.880]		[1.000]	[.998]
	$N = 4,782$	$N = 4,782$	$N = 4,712$	$N = 4,782$	$N = 4,782$	$N = 4,712$
Management priority on health/safety [survey]	0.790	0.057***	0.050***	0.784	0.028*	0.024
		(0.015)	(0.014)		(0.016)	(0.015)
		[.001]	[.003]		[.539]	[.632]
	$N = 3,566$	$N = 3,566$	$N = 3,514$	$N = 3,018$	$N = 3,018$	$N = 2,976$
Productivity index [survey/admin]	0.000	−0.046	−0.060	0.000	−0.015	−0.054
		(0.061)	(0.056)		(0.062)	(0.056)
		[.450]	[.283]		[.805]	[.328]
	$N = 3,309$	$N = 3,309$	$N = 3,300$	$N = 2,890$	$N = 2,890$	$N = 2,881$

TABLE III
(CONTINUED)

Outcome variable	First year (12 months)			Longer-run (24–30 months)		
	Mean (1)	No controls (2)	Post-Lasso (3)	Mean (4)	No controls (5)	Post-Lasso (6)
Panel C: Health status and behaviors						
IL Marathon/10K/5K [admin]	0.066	0.002	−0.005	0.052	0.006	0.001
		(0.008)	(0.006)		(0.007)	(0.006)
		[.975]	[.471]		[.625]	[.995]
	$N = 4{,}834$	$N = 4{,}834$	$N = 4{,}817$	$N = 4{,}834$	$N = 4{,}834$	$N = 4{,}817$
Campus gym visits (days/year) [admin]	5.839	−0.062	0.401	5.047	−0.342	0.001
		(0.733)	(0.360)		(0.660)	(0.367)
		[.975]	[.471]		[.625]	[.998]
	$N = 4{,}834$	$N = 4{,}834$	$N = 4{,}817$	$N = 4{,}834$	$N = 4{,}834$	$N = 4{,}817$
Ever screened [survey]	0.942	0.039***	0.036***	0.962	0.029***	0.027***
		(0.009)	(0.008)		(0.008)	(0.007)
		[.001]	[.000]		[.006]	[.005]
	$N = 3{,}567$	$N = 3{,}567$	$N = 3{,}557$	$N = 3{,}020$	$N = 3{,}020$	$N = 3{,}010$

Notes. Each estimate is from a separate regression of an outcome, specified by the row, on a treatment group indicator. Observations include the control and treatment groups. Longer-run time horizons are 24 and 30 months for survey and admin outcomes, respectively. Post-Lasso specifications control for covariates selected by Lasso to predict the outcome. Potential predictors include all available baseline variables in the same family of outcomes, strata variables, and the baseline survey variables reported in Table I, as well as two-way interactions between these predictors. Robust standard errors are reported in parentheses. */**/*** indicates significance at the 10%/5%/1% level using conventional inference. Family-wise *p*-values, reported in brackets, adjust for the number of outcome variables in each family. Results for all outcomes, categorized by family, are reported in Online Appendix Tables A.4a–A.4g (12-month outcomes) and Tables A.7a–A.7g (longer-run outcomes).

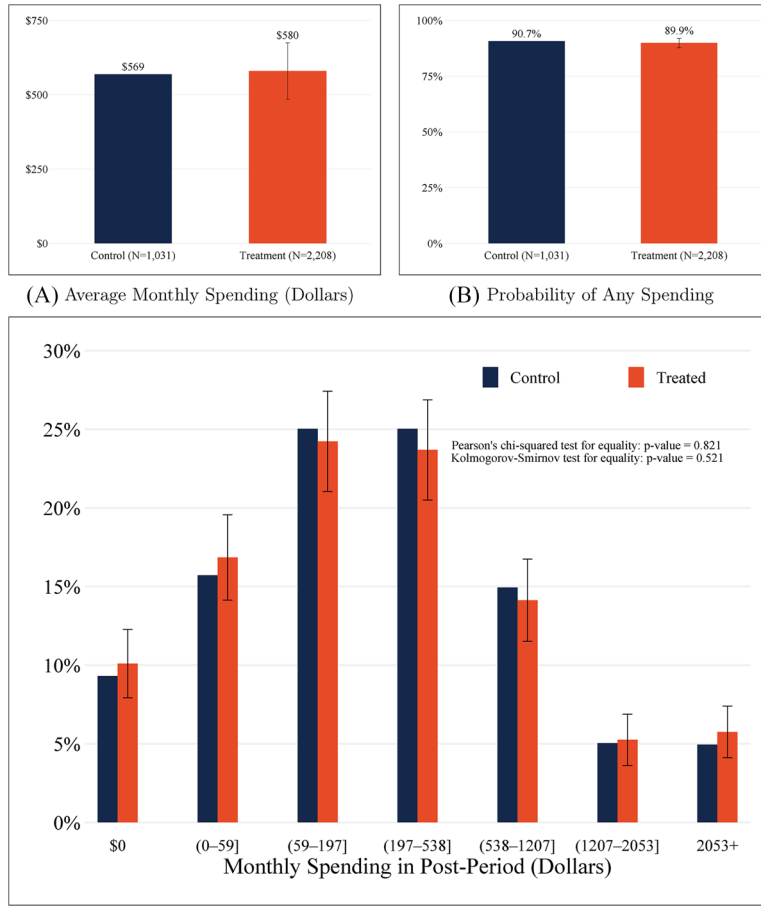
survey outcomes is provided in Online Appendix Tables A.4a–A.4g.

We report ITT estimates using two specifications. The first includes no control variables, and the second specification includes a set of baseline outcomes and covariates chosen via Lasso, as described in Section III.B. Because the probability of treatment assignment was constant across strata, these controls are included not to reduce bias but to improve the precision of the treatment effect estimates (Bruhn and McKenzie 2009). For completeness, the appendix tables also report a third control specification that includes fixed effects for the 69 strata used for stratified random assignment at baseline.

2. Medical Spending. We do not detect statistically significant effects of treatment on average medical spending over the first 12 months (August 2016–July 2017) of the wellness intervention in any of our specifications. Table III, column (2), first row shows that average monthly spending was $10.8 higher in the treatment group than in the control group. The point estimate increases slightly when using the post-Lasso control strategy (column (3)) but remains small and statistically indistinguishable from 0. The post-Lasso specification improves the estimate's precision, with a standard error about 24% smaller than that of the no-control specification. Panel A, columns (2)–(3) also show small and insignificant effects for different subcategories of spending and the probability of any spending over this 12-month period.

Figure V, Panels A and B graphically reproduce the null average treatment effects presented in Table III, Panel A, column (2) for total and nonzero spending. Despite null effects on average, there may still exist mean-preserving treatment effects that alter other moments of the spending distribution. However, Figure V, Panel C shows that the empirical distributions of spending are observationally similar for both the treatment and control groups. This similarity is formalized by a Pearson's chi-squared test and a Kolmogorov-Smirnov test, which both fail to reject the null hypothesis that the control and treatment samples were drawn from the same spending distribution ($p = .821$ and $p = .521$, respectively).

3. Employment and Productivity. Next we estimate the effect of treatment on various employment and productivity outcomes. Table III, Panel B, columns (2) and (3) summarize our



(A) Average Monthly Spending (Dollars)

(B) Probability of Any Spending

(C) Histogram of Average Monthly Spending, by Quantile of Control Group Spending ($N = 3{,}239$)

FIGURE V

Postintervention Medical Spending by Treatment Status

Results based on health care claims over the 12-month period August 2016–July 2017. The null hypothesis of the Pearson's chi-squared and the nonparametric Kolmogorov-Smirnov tests is that the two samples are drawn from the same distribution.

findings, while Online Appendix Tables A.4c and A.4d report estimates for all administrative and prespecified survey productivity measures. We do not detect statistically significant effects after 12 months of the wellness intervention on any of our administratively measured outcomes, including annual salary,

the probability of job promotion or job termination, and sick leave taken.

Among self-reported employment and productivity outcomes measured by the one-year follow-up survey, we find no statistically significant effects on most measures, including being happier at work than last year or feeling very productive at work. The only exception is that individuals in the treatment group are 5.7 percentage points (7.2%) more likely (family wise $p = .001$) to believe that management places a priority on health and safety (Table III, column (2)). The treatment effect on the 12-month productivity index, equal to the first principal component of all 12-month survey and administrative employment and productivity outcomes, is statistically insignificant.

Table III, Panel B, column (1) reports that 17.6% of our sample had received a promotion and 11.3% had ceased employment by the end of the first year, suggesting that our null estimates are not due to stickiness in career progression.[19] A more serious concern is whether our productivity measures are sufficiently meaningful and/or precise to draw conclusions. Following Baker, Gibbs, and Holmström (1994), we cross-validate our administrative measures of employment and productivity, comparing each to our survey measures of work and productivity. As reported in Online Appendix Table A.9, we find a strong degree of concordance between the independently measured administrative and survey variables. The eighth row of column (3) reports that individuals who self-report receiving "a promotion or more responsibility at work" are 22.5% more likely to have an official title change in our administrative data, and column (2) reports that they are 22.9% more likely to have received a promotion, which we define as having both a job title change and a nonzero salary raise.[20]

More generally, our administrative measure of promotion is positively correlated with self-reported job satisfaction and happiness at work and negatively correlated with self-reported job search. Likewise, the first row of column (5) reports that survey

19. There is even less stickiness in the longer-run estimates reported in Section V, where our precision allows us to reject small increases in productivity during the first 30 months following randomization.

20. As discussed in Section II.C, less than 5% of employees with job title changes did not also have a salary raise. We obtain a similar causal effect estimate if we look only at job title changes rather than our constructed promotion measure (see Online Appendix Table A.4c).

respondents who indicated they had taken any sick days were recorded in the administrative data as taking 3.2 more sick days than respondents who had not indicated taking sick days. The high overall agreement between our survey and administrative variables both increases our confidence in their accuracy and validates their relevance as measures of productivity.

4. Health Behaviors. Finally, we investigate health behaviors, which may respond more quickly to a wellness intervention than medical spending or productivity. Our main results are reported in Table III, Panel C, columns (2) and (3). We find small and statistically insignificant treatment effects on participation in any running event of the April 2017 Illinois Marathon (5K, 10K, or half/full marathons). Similarly, we do not find meaningful effects on the average number of days a month that an employee visits a campus recreation facility. However, we do find that individuals in the treatment group are nearly 4 percentage points more likely (family wise $p = .001$) to report ever having a previous health screening. This effect indicates that our intervention's biometric health screenings did not simply crowd out screenings that would have otherwise occurred within the first year of our study.

5. Discussion. Across all 42 outcomes we examine, we find only two statistically significant effects of our intervention after one year: an increase in the number of employees who ever received a health screening and an increase in the number who believe that management places a priority on health and safety.[21] The next section addresses the precision of our estimates by quantifying what effects we can rule out, but first we mention a few caveats.

First, these results only include one year of data. Although we do not find significant effects for most of the outcomes we examine, it is possible that longer-run effects may emerge in later years, so we turn to this issue in Section V. Second, our analysis assumes that the control group was unaffected by the intervention. The research team's contact with the control group in the first year was confined to the communication procedures employed for the 2016 and 2017 online surveys.

21. We show in the Online Appendix that these two effects are driven by the health screening component of our intervention rather than the wellness activity component.

Although we never shared details of the intervention with members of the control group, they may have learned or have been affected by the intervention through peer effects. However, we think peer effects are unlikely to explain our null findings. We asked study participants on the 2017 follow-up survey whether they ever talked about the iThrive workplace wellness program with any of their coworkers. Only 3% of the control group responded affirmatively, compared with 44% of the treatment group. Moreover, the cluster-randomized trial of Song and Baicker (2019), which has a design that naturally accommodates peer effects, also finds null effects of a comprehensive workplace wellness program.

Finally, our results do not rule out the possibility of meaningful treatment effect heterogeneity. There may exist subpopulations who did benefit from the intervention or who would have benefited had they participated. Wellness programs vary considerably across employers, and another design that induces a different population to participate, such as by forgoing a biometric screening, may achieve different results from what we find here.

6. Comparison with Prior Studies. We now compare our estimates to the prior literature, which has focused on medical spending and absenteeism. This exercise employs a spending estimate derived from a data sample that winsorizes (top codes) medical spending at the one percent level (see Online Appendix Table A.11, column (3)). We do this to reduce the influence of a small number of extreme outliers on the precision of our estimate, as in prior studies (e.g., Clemens and Gottlieb 2014).[22]

Figure VI illustrates how our estimates compare to the prior literature.[23] The top-left figure in Panel A plots the distribution of the ITT point estimates for medical spending from 22 prior workplace wellness studies. The figure also plots our ITT point estimate for total medical spending from Table III and shows that our 95% confidence interval rules out 20 of these 22 estimates. For ease

22. Winsorizing can introduce bias if there are heterogeneous treatment effects in the tails of the spending distribution. However, Figure V, Panel C provides evidence of a consistently null treatment effect throughout the spending distribution. This evidence is further supported by Online Appendix Table A.11, which shows that the point estimate of the medical spending treatment effect changes little after winsorization. For completeness, Online Appendix Figure A.1 illustrates the stability of the point estimate across a wide range of winsorization levels.

23. Online Appendix B provides the sources and calculations underlying the point estimates reported in Figure VI.

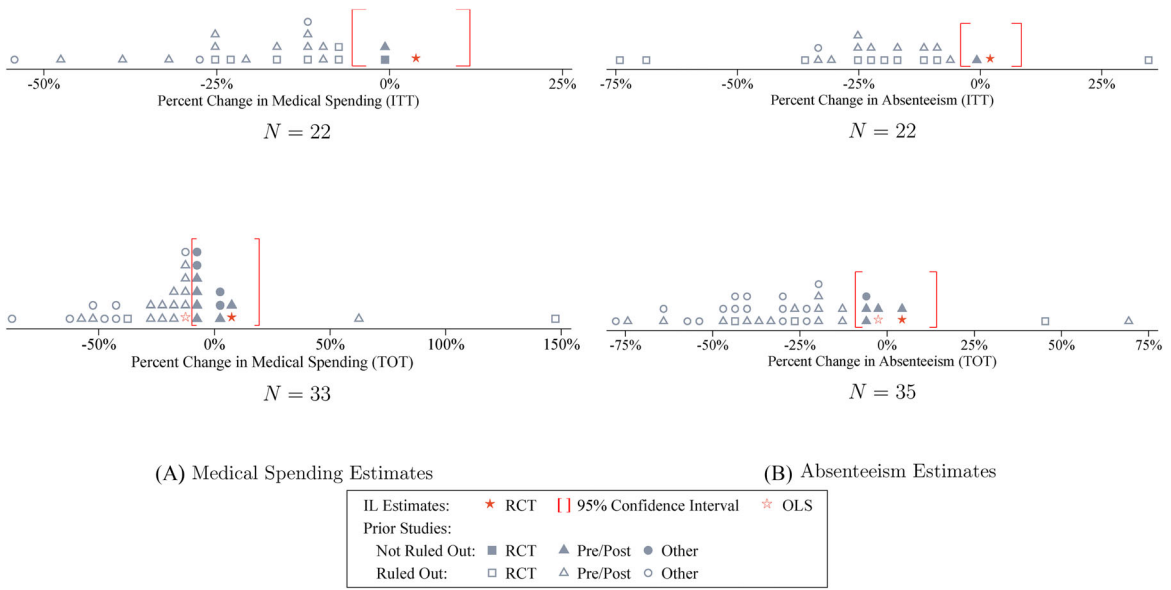


FIGURE VI

Comparison of Experimental Estimates to Prior Studies

Each panel shows the distribution of N point estimates from prior workplace wellness studies. Panel A plots intent-to-treat (ITT) and treatment-on-the-treated (TOT) estimates for medical spending. Panel B plots corresponding estimates for absenteeism. The point estimates from our own study (RCT Estimate), and their associated confidence intervals, are taken from Online Appendix Table A.11, column (3) for medical spending and Table III, column (3) and Table IV, column (2) for absenteeism. Our randomized controled trial (RCT) estimates and confidence intervals are plotted to demonstrate the share of prior study point estimates we can rule out. Online Appendix Table B.1 provides the full details of this meta-analysis.

of comparison, all effects are expressed as percent changes. The bottom-left figure in Panel A plots the distribution of treatment-on-the-treated (TOT) estimates for health spending from 33 prior studies, along with the IV estimates from our study. In this case, our 95% confidence interval rules out 23 of the 33 studies. Overall, our confidence intervals rule out 43 of 55 (78%) prior ITT and TOT point estimates for health spending.[24]

The two figures in Panel B repeat this exercise for absenteeism and show that our estimates rule out 51 of 57 (89%) prior ITT and TOT point estimates for absenteeism. Across both sets of outcomes, we rule out 94 of 112 (84%) prior estimates. If we restrict our comparison to the studies that lasted 12 months or less, we rule out 39 of 47 (83%) prior estimates, and if we restrict our comparison to only the set of RCTs, we rule out 21 of 22 (95%) prior estimates. If we combine RCTs and studies that use a pre- or postdesign, we continue to rule out 68 of 81 (84%) prior estimates.

We can also combine our spending and absenteeism estimates with our cost data to calculate an ROI for workplace wellness programs. The 99% confidence interval for the ROI associated with our intervention rules out the widely cited savings estimates reported in the meta-analysis of Baicker, Cutler, and Song (2010).[25] One reason for the divergence between our estimates and prior findings may be selection bias in observational studies, which we explore in Section IV.C. However, our estimates differ even when we restrict comparisons to prior RCTs. Another possible explanation in these cases is publication bias. Using the method of Andrews and Kasy (2019) on the subset of prior studies that report standard errors ($N = 40$), our results in Online Appendix Table A.13 suggest that the bias-corrected mean effect in these studies is negative but insignificant ($p = .14$). Furthermore, studies with p-values greater than .05 appear to be only one-third as

24. If we do not winsorize medical spending, we rule out 40 of 55 (73%) prior health studies.

25. The first year of the iThrive program cost \$152 (= \$271 × 0.56) per person assigned to treatment. This is a conservative estimate because it does not account for paid time off or the fixed costs of managing iThrive. Focusing on the first year of our intervention and assuming that the cost of a sick day equals \$240, we calculate that the lower bounds of the 99% confidence intervals for annual medical and absenteeism costs are −\$396 (=(17.2 − 2.577 × 19.5) × 12) and −\$91 (=(0.138 − 2.577 × 0.200) × 240), which imply ROI lower bounds of 2.61 and 0.60, respectively. By comparison, Baicker, Cutler, and Song (2010) found that spending fell by \$3.27, and absenteeism costs by \$2.73, for every dollar spent on wellness programs.

likely to be published as studies with significantly negative effects on spending and absenteeism.

7. IV versus OLS. As shown above, our results differ from many prior studies that find workplace wellness programs significantly reduce health expenditures and absenteeism. One possible reason for this discrepancy is that our results may not generalize to other workplace populations or programs. A second possibility is the presence of advantageous selection bias in these other studies, which are generally not RCTs (Oster 2019). We investigate the potential for selection bias to explain this difference by performing a typical observational (OLS) analysis and comparing its results to those of our experimental estimates.[26] Specifically, we estimate

$$Y_i = \alpha + \gamma P_i + \Gamma X_i + \varepsilon_i, \tag{3}$$

where Y_i is the outcome variable as in equation (2), P_i is an indicator for participating in the screening and HRA, and X_i is a vector of variables that control for potentially nonrandom selection into participation.

We estimate two variants of equation (3). The first is an IV specification that includes observations for individuals in the treatment or control groups and uses treatment assignment as an instrument for completing the first-year screening and HRA. The second variant estimates equation (3) using OLS, restricted to individuals in the treatment group. For these two variants, we estimate three specifications similar to those used for the ITT analysis described above (no controls, strata fixed effects, and post-Lasso).[27] This generates six estimates for each outcome variable. Table IV reports the "no controls" and "post-Lasso" results for our primary outcomes of interest. Results for all specifications,

26. This observational analysis was not specified in our preanalysis plan.

27. To select controls for the post-Lasso IV specification, we follow the "triple" selection strategy proposed in Chernozhukov, Hansen, and Spindler (2015). This strategy first estimates three Lasso regressions of (i) the (endogenous) focal independent variable on all potential controls and instruments, (ii) the focal independent variable on all potential controls, and (iii) the outcome on all potential controls. It then forms a 2SLS estimator using instruments selected in step i and all controls selected in any of the steps i–iii. When the instrument is randomly assigned, as in our setting, the set of controls selected in steps i–ii will be the same, in expectation. Thus, we form our 2SLS estimator using treatment assignment as the instrument and controls selected in Lasso steps ii or iii of this algorithm.

TABLE IV

FIRST-YEAR TREATMENT EFFECTS: EXPERIMENTAL VERSUS OBSERVATIONAL ESTIMATES

Outcome variable	Experimental (IV)		Observational (OLS)	
	No controls (1)	Post-Lasso (2)	No controls (3)	Post-Lasso (4)
Panel A: Medical spending				
Total spending (dollars/month) [admin]	17.7	52.3	−137.3**	−103.8*
	(79.0)	(59.4)	(68.6)	(61.9)
	$N = 3,239$	$N = 3,152$	$N = 2,208$	$N = 2,140$
Drug spending [admin]	−13.8	−12.8	−26.3	−7.3
	(43.2)	(20.4)	(27.2)	(12.0)
	$N = 3,239$	$N = 3,152$	$N = 2,208$	$N = 2,140$
Office spending [admin]	−9.9	−3.1	12.2	8.7*
	(16.2)	(6.8)	(7.5)	(5.1)
	$N = 3,239$	$N = 3,152$	$N = 2,208$	$N = 2,140$
Hospital spending [admin]	36.1	45.2	−118.0**	−83.4
	(50.4)	(45.6)	(55.7)	(51.8)
	$N = 3,239$	$N = 3,152$	$N = 2,208$	$N = 2,140$
Nonzero medical spending [admin]	−0.013	0.004	0.061***	0.036***
	(0.018)	(0.016)	(0.014)	(0.012)
	$N = 3,239$	$N = 3,152$	$N = 2,208$	$N = 2,140$
Panel B: Employment and productivity				
Job promotion [admin]	−0.006	−0.009	0.019	0.009
	(0.022)	(0.021)	(0.014)	(0.015)
	$N = 4,146$	$N = 4,130$	$N = 2,840$	$N = 2,828$
Job terminated [admin]	−0.022	−0.023	−0.080***	−0.063***
	(0.018)	(0.017)	(0.011)	(0.011)
	$N = 4,834$	$N = 4,753$	$N = 3,300$	$N = 3,244$
Sick leave (days/year) [admin]	0.322	0.224	0.275	−0.068
	(0.398)	(0.344)	(0.272)	(0.251)
	$N = 4,782$	$N = 4,712$	$N = 3,264$	$N = 3,216$
Management priority on health/safety [survey]	0.087***	0.077***	−0.004	−0.007
	(0.023)	(0.021)	(0.017)	(0.016)
	$N = 3,566$	$N = 3,514$	$N = 2,410$	$N = 2,376$
Productivity index [survey/admin]	−0.070	−0.096	0.069	0.083
	(0.092)	(0.085)	(0.073)	(0.067)
	$N = 3,309$	$N = 3,300$	$N = 2,245$	$N = 2,240$
Panel C: Health status and behaviors				
IL Marathon/10K/5K 2017 [admin]	0.003	−0.011	0.059***	0.024***
	(0.014)	(0.011)	(0.008)	(0.006)
	$N = 4,834$	$N = 4,817$	$N = 3,300$	$N = 3,287$
Campus gym visits (days/year) [admin]	−0.110	0.757	3.527***	2.160***
	(1.309)	(0.656)	(0.813)	(0.425)
	$N = 4,834$	$N = 4,817$	$N = 3,300$	$N = 3,287$
Ever screened [survey]	0.060***	0.056***	0.073***	0.061***
	(0.014)	(0.012)	(0.011)	(0.009)
	$N = 3,567$	$N = 3,557$	$N = 2,410$	$N = 2,404$

Notes. Each row and column reports estimates from a separate regression. The outcome in each regression is specified by the table row, and the (endogenous) focal independent variable is an indicator for completing the screening and health risk assessments (HRAs). For the IV specifications (columns (1)–(2)), the instrument is an indicator for inclusion in the treatment group, and observations include individuals in the control or treatment groups. For the OLS specifications (columns (3)–(4)), there is no instrument and observations are restricted to individuals in the treatment group. The control strategy is specified by the column. Post-Lasso controls include covariates selected by Lasso to predict either the dependent variable or the focal independent variable. The set of potential predictors include baseline values of all available variables in the same family of outcomes, strata variables, and the baseline (2016) survey variables reported in Table I, as well as all two-way interactions between these predictors. Robust standard errors are reported in parentheses. */**/*** indicates significance at the 10%/5%/1% level using conventional inference.

including strata fixed effects, and all prespecified administrative and survey outcomes are reported in Online Appendix Tables A.5a–A.5h. Comparing OLS estimates to IV estimates for the post-Lasso specification, which chooses controls from a large set of variables, illustrates the extent to which rich controls can mitigate selection bias in an observational analysis.

As with the ITT analysis, the IV estimates reported in Table IV, columns (1) and (2) are small and indistinguishable from 0 for nearly every outcome. By contrast, the observational estimates reported in columns (3) and (4) are frequently large and statistically significant. Moreover, the IV estimate rules out the OLS estimate for several outcomes. Based on our most precise and well-controlled specification (post-Lasso), the OLS monthly spending estimate of −$103.8 (row 1, column (4)) lies outside the 99% confidence interval of the IV estimate of $52.3 with a standard error of $59.4 (row 1, column (2)). For participation in the April 2017 Illinois Marathon/10K/5K, the OLS estimate of 0.024 lies outside the 99% confidence interval of the corresponding IV estimate of −0.011. For campus gym visits, the OLS estimate of 2.160 lies just inside the 95% confidence interval of the corresponding IV estimate of 0.757. Under the assumption that the IV (RCT) estimates are asymptotically consistent, these differences imply that even after conditioning on a rich set of controls, participants selected into our workplace wellness program on the basis of lower-than-average contemporaneous spending and healthier-than-average behaviors. This selection bias is consistent with the evidence presented in Section III.A that preexisting spending is lower, and preexisting behaviors are healthier, among participants than among nonparticipants.

Moreover, the observational estimates presented in columns (3)–(4) are in line with estimates from previous observational studies, which suggests that our setting is not particularly unique. In the spirit of LaLonde (1986), these estimates demonstrate that even well-controlled observational analyses can suffer from significant selection bias, suggesting that similar biases are present in other wellness program settings as well.

V. Longer-Run Results

The first year of our intervention concluded in July 2017. We continued to offer the iThrive wellness program to the treatment group for a second year (August 2017–July 2018). We maintained

the same basic structure as in the first year but offered smaller incentives—a design choice influenced both by a smaller budget and the diminishing effect of incentives on participation that we observed during the first year.[28] In particular, the second year of iThrive again included a health screening, an HRA, and a set of wellness activities offered in both the fall and spring semesters. iThrive officially ended in September 2018 with a third and final health screening.

This section reports estimates of the causal, ITT effect of our two-year intervention on longer-run outcomes using data that extend up to two and a half years (30 months) postrandomization. We note that our study design entailed offering follow-up health screenings to the treatment and control groups in 2017 and 2018, one and two years after the intervention began, respectively. This means the control group received a partial treatment, which potentially attenuates treatment effect estimates beyond 12 months for outcomes affected by screening in the short run. However, the scope for attenuation is limited. Control group participants were eligible only to receive a health screening; they were ineligible for both the HRA and the wellness activities. Moreover, we know from our estimates above that even the full intervention—screening, HRA, and wellness activities—had little effect on most outcomes during the first 12 months.

Table III, columns (5) and (6) summarize our primary treatment effect estimates after 24 months for survey outcomes and 30 months for admin outcomes (time horizons based on data availability).[29] Overall, the longer-run estimates are qualitatively similar to those from the one-year analysis. Notably, we continue to find no effects on job promotion despite a mean 30-month promotion rate of 36%. The 30-month effect on job termination, which at 12 months was insignificant at −1.2 percentage points, is now very close to 0 (0.2 percentage points) despite a mean 30-month termination rate of 20.4%. Our 95% confidence interval for job termination rules out a positive retention effect of 2.4 percentage points (12%) for iThrive. For perspective, this upper bound is well below the 4.3 percentage points needed to generate the screening savings discussed in Section IV.B.

28. Online Appendix Figure D.2 illustrates the structure of incentives and treatments offered in the second year of the wellness program.

29. Longer-run results for all outcomes and control specifications are shown in Online Appendix Tables A.7a–A.7g.

Although we previously found that individuals in the treatment group were more likely to believe management places a priority on health and safety after the first year, the two-year estimate is attenuated and is no longer statistically significant in our preferred (post-Lasso) specification. We continue to find that individuals in the treatment group are more likely to report having a previous health screening, and this effect remains statistically significant (family wise $p = .005$).

The point estimate for 30-month total medical spending is lower than the first-year estimates, and the standard error has increased. The reduction in precision is likely caused by outliers, as described previously in Section IV.C. As with our 12-month estimates, we reduce the influence of outliers by winsorizing at the 1% level. Spending estimates at various levels of winsorization are presented in Online Appendix Table A.12. For 1% winsorization (column (3)), we estimate an ITT effect of \$5.7 with a 95% confidence interval of [−33.8, 45.1]. This is very similar to the winsorized 12-month estimate of \$17.2 and 95% confidence interval of [−21.0, 55.3] (Online Appendix Table A.11, column (3)).

Increasing the length of the follow-up window raises concerns about the potential for differential attrition between the control and treatment groups. However, Online Appendix Table A.10 shows that health insurance enrollment is nearly identical in the control and treatment groups over both the 12- and 30-month postperiods. In addition, the rates of job exit, which measure sample attrition for outcomes derived from university administrative data and the rates of completion for the one-year follow-up survey, are also similar. We do detect a small but statistically significant difference in completion rates for the second year (2018) follow-up survey. The completion rates remain fairly high for both the treatment and control groups, but the difference in completion suggests that outcomes derived from the two-year follow-up survey should potentially be weighted less than those from other data sources.

VI. Conclusion

This article evaluates a two-year comprehensive workplace wellness program, iThrive, that we designed and implemented. We find that employees who chose to participate in our wellness program were less likely to be in the bottom quartile of the income distribution and already had lower medical spending and healthier behaviors than nonparticipants prior to our

intervention. These selection effects imply that workplace wellness programs may shift costs onto low-income employees with high health care spending and poor health habits. Moreover, the large magnitude of our selection on prior spending suggests that a potential value of wellness programs to firms may be their potential to attract and retain workers with low health care costs.

The iThrive wellness program increased lifetime health screening rates but had no effects on medical spending, health behaviors, or employee productivity after 30 months. Our null results are economically meaningful: we can rule out 84% of the medical spending and absenteeism estimates from the prior literature along with the average ROIs calculated by Baicker, Cutler, and Song (2010) in a widely cited meta-analysis. Our OLS estimate is consistent with results from the prior literature, but was ruled out by our IV estimate, suggesting that non-RCT studies in this literature suffer from selection bias.

Well-designed studies have found that monetary incentives can successfully promote exercise (e.g., Charness and Gneezy 2009), and there is ample evidence that exercise improves health (e.g., Warburton et al. 2006). However, both our 30-month study and the Song and Baicker (2019) 18-month study find null effects of workplace wellness on primary outcomes of interest despite using different program and randomization designs and examining different populations. These null findings underscore the challenges to achieving health benefits with large-scale wellness interventions, a point echoed by Cawley and Price (2013). One potential explanation for these disappointing results could be that those who benefit the most (e.g., smokers and those with high medical costs) decline to participate, even when offered large monetary incentives. An improved understanding of participation decisions would help wellness programs better target these individuals.

UNIVERSITY OF CHICAGO AND NBER
UNIVERSITY OF ILLINOIS AND NBER
UNIVERSITY OF ILLINOIS AND NBER

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online. Code replicating tables and figures in this article can be found in Jones, Molitor, and Reif (2019), in the Harvard Dataverse, doi:10.7910/DVN/VELJKG.

REFERENCES

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," NBER Working Paper no. 24003, 2017.

Abraham, Jean, and Katie M. White, "Tracking the Changing Landscape of Corporate Wellness Companies," *Health Affairs*, 36 (2017), 222–228.

Aldana, Steven G., "Financial Impact of Health Promotion Programs: A Comprehensive Review of the Literature," *American Journal of Health Promotion*, 15 (2001), 296–320.

Andrews, Isaiah, and Maximilian Kasy, "Identification of and Correction for Publication Bias," *American Economic Review*, 109 (2019), 2766–2794.

Baicker, Katherine, David Cutler, and Zirui Song, "Workplace Wellness Programs Can Generate Savings," *Health Affairs*, 29 (2010), 304–311.

Baker, George, Michael Gibbs, and Bengt Holmström, "The Internal Economics of the Firm: Evidence from Personnel Data," *Quarterly Journal of Economics*, 109 (1994), 881–919.

Baxter, Siyan, Kristy Sanderson, Alison J. Venn, C. Leigh Blizzard, and Andrew J. Palmer, "The Relationship between Return on Investment and Quality of Study Methodology in Workplace Health Promotion Programs," *American Journal of Health Promotion*, 28 (2014), 347–363.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81 (2014), 608–650.

Bruhn, Miriam, and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1 (2009), 200–232.

Burd, Steven A., "How Safeway Is Cutting Health-Care Costs," *Wall Street Journal*, June 12, 2009, http://www.wsj.com/articles/SB124476804026308603.

Cawley, John, "The Affordable Care Act Permits Greater Financial Rewards for Weight Loss: A Good Idea in Principle, but Many Practical Concerns Remain," *Journal of Policy Analysis and Management*, 33 (2014), 810–820.

Cawley, John, and Joshua A. Price, "A Case Study of a Workplace Wellness Program That Offers Financial Incentives for Weight Loss," *Journal of Health Economics*, 32 (2013), 794–803.

Chapman, Larry S., "Meta-Evaluation of Worksite Health Promotion Economic Return Studies: 2012 Update," *American Journal of Health Promotion*, 26 (2012), 1–12.

Charness, Gary, and Uri Gneezy, "Incentives to Exercise," *Econometrica*, 77 (2009), 909–931.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler, "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," *American Economic Review*, 105 (2015), 486–490.

Clemens, Jeffrey, and Joshua D. Gottlieb, "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?," *American Economic Review*, 104 (2014), 1320–1349.

Gowrisankaran, Gautam, Karen Norberg, Steven Kymes, Michael E. Chernew, Dustin Stwalley, Leah Kemper, and William Peck, "A Hospital System's Wellness Program Linked to Health Plan Enrollment Cut Hospitalizations but Not Overall Costs," *Health Affairs*, 32 (2013), 477–485.

Gubler, Timothy, Ian Larkin, and Lamar Pierce, "Doing Well by Making Well: The Impact of Corporate Wellness Programs on Employee Productivity," *Management Science*, 64 (2017), 4967–5460.

Haisley, Emily, Kevin G. Volpp, Thomas Pellathy, and George Loewenstein, "The Impact of Alternative Incentive Schemes on Completion of Health Risk Assessments," *American Journal of Health Promotion*, 26 (2012), 184–188.

Handel, Benjamin, and Jonathan Kolstad, "Wearable Technologies and Health Behaviors: New Data and New Methods to Understand Population

Health," *American Economic Review: Papers and Proceedings*, 107 (2017), 481–485.
Horwitz, Jill R., Brenna D. Kelly, and John E. DiNardo, "Wellness Incentives in the Workplace: Cost Savings through Cost Shifting to Unhealthy Workers," *Health Affairs*, 32 (2013), 468–476.
Jaspen, Bruce, "Employers Boost Wellness Spending 17% from Yoga to Risk Assessments," *Forbes Online*, March 26, 2015, http://www.forbes.com/sites/brucejapsen/2015/03/26/employers-boost-wellness-spending-17-from-yoga-to-risk-assessments/#6a37ebf2350f.
Jones, Damon, David Molitor, and Julian Reif, "Replication Data for: 'What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study'," (2019), Harvard Dataverse, doi: 10.7910/DVN/VELJKG.
Kaiser Family Foundation, "Workplace Wellness Programs Characteristics and Requirements," Kaiser Family Foundation, 2016, http://files.kff.org/attachment/Issue-Brief-Workplace-Wellness-Programs-Characteristics-and-Requirements.
Kaiser Family Foundation and Health Research & Educational Trust, "Employer Health Benefits: 2016 Annual Survey," Kaiser Family Foundation, 2016a. http://files.kff.org/attachment/Report-Employer-Health-Benefits-2016-Annual-Survey.
———, "Employer Health Benefits: 2017 Annual Survey," Kaiser Family Foundation, 2017, http://files.kff.org/attachment/Report-Employer-Health-Benefits-Annual-Survey-2017.
LaLonde, Robert J., "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76 (1986), 604–620.
Lazear, Edward P., "Performance Pay and Productivity," *American Economic Review*, 90 (2000), 1346–1361.
Liu, Tim, Christos Makridis, Paige Ouimet, and Elena Simintzi, "Is Cash Still King: Why Firms Offer Non-Wage Compensation and the Implications for Shareholder Value," University of North Carolina at Chapel Hill Working Paper, 2017, https://ssrn.com/abstract=3088067.
Mattke, Soeren, Harry H. Liu, John P. Caloyeras, Christina Y. Huang, Kristin R. Van Busum, Dmitry Khodyakov, and Victoria Shier, "Workplace Wellness Programs Study: Final Report," RAND Corporation, 2013.
Mattke, Soeren, Christopher Schnyer, and Kristin R. Van Busum, "A Review of the U.S. Workplace Wellness Market," RAND Corporation, 2012, Occasional Paper Series, https://www.dol.gov/sites/default/files/ebsa/researchers/analysis/health-and-welfare/workplacewellnessmarketreview2012.pdf.
McIntyre, Adrianna, Nicholas Bagley, Austin Frakt, and Aaron Carroll, "The Dubious Empirical and Legal Foundations of Workplace Wellness Programs," *Health Matrix*, 27 (2017), 59.
Meenan, Richard T., Thomas M. Vogt, Andrew E. Williams, Victor J. Stevens, Cheryl L. Albright, and Claudio Nigg, "Economic Evaluation of a Worksite Obesity Prevention and Intervention Trial among Hotel Workers in Hawaii," *Journal of Occupational and Environmental Medicine/American College of Occupational and Environmental Medicine*, 52 (2010), S8.
Oster, Emily, "Behavioral Feedback: Do Individual Choices Influence Scientific Results?" NBER Working Paper no. w25225, 2018.
Oyer, Paul, "Salary or Benefits?" In *Work, Earnings and Other Aspects of the Employment Relation*, Solomon W. Polachek and Konstantinos Tatsiramos, eds. (Bingley, UK: JAI Press, 2008), 429–467.
Pelletier, Kenneth R., "A Review and Analysis of the Clinical and Cost-Effectiveness Studies of Comprehensive Health Promotion and Disease Management Programs at the Worksite: Update VIII 2008 to 2010," *Journal of Occupational and Environmental Medicine*, 53 (2011), 1310–1331.
Reynolds, Chelsea, "Myth Surrounds Reform's 'Safeway Amendment'," *Covering Health*, January 20, 2010, http://healthjournalism.org/blog/2010/01/myth-surrounds-reforms-safeway-amendment/.

Royer, Heather, Mark Stehr, and Justin Sydnor, "Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company," *American Economic Journal: Applied Economics*, 7 (2015), 51–84.

Ryde, Gemma C., Nicholas D. Gilson, Nicola W. Burton, and Wendy J. Brown, "Recruitment Rates in Workplace Physical Activity Interventions: Characteristics for Success," *American Journal of Health Promotion*, 27 (2013), e101–e112.

Saunders, Rob, Madhu Vulimiri, Mark Japinga, William Bleser, and Charlene Wong, "Are Carrots Good for Your Health? Current Evidence on Health Behavior Incentives in the Medicaid Program," Duke Margolis Center for Health Policy, 2018, https://healthpolicy.duke.edu/sites/default/files/atoms/files/duke_healthybehaviorincentives_6.1.pdf.

Soler, Robin E., Kimberly D. Leeks, Sima Razi, David P. Hopkins, Matt Griffith, Adam Aten, Sajal K. Chattopadhyay, Susan C. Smith, Nancy Habarta, Ron Z. Goetzel, Nicolaas P. Pronk, Dennis E. Richling, Deborah R. Bauer, Leigh Ramsey Buchanan, Curtis S. Florence, Lisa Koonin, Debbie MacLean, Abby Rosenthal, Dyann Matson Koffman, James V. Grizzell, and Andrew M. Walker, "A Systematic Review of Selected Interventions for Worksite Health Promotion: The Assessment of Health Risks with Feedback," *American Journal of Preventive Medicine*, 38 (2010), S237–S262.

Song, Zirui, and Katherine Baicker, "Effect of a Workplace Wellness Program on Employee Health and Economic Outcomes: A Randomized Clinical Trial," *Journal of the American Medical Association*, 321 (2019), 1491–1501.

Terry, Paul E., Jinnet Briggs Fowles, Min Xi, and Lisa Harvey, "The ACTIVATE Study: Results from a Group-Randomized Controlled Trial Comparing a Traditional Worksite Health Promotion Program with an Activated Consumer Program," *American Journal of Health Promotion*, 26 (2011), e64–e73.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov, "Using Double-Lasso Regression for Principled Variable Selection," University of Chicago Working Paper, 2016.

Volpp, Kevin G., David A. Asch, Robert Galvin, and George Loewenstein, "Redesigning Employee Health Incentives: Lessons from Behavioral Economics," *New England Journal of Medicine*, 365 (2011), 388–390.

Volpp, Kevin G., Leslie K. John, Andrea B. Troxel, Laurie Norton, Jennifer Fassbender, and George Loewenstein, "Financial Incentive-based Approaches for Weight Loss: A Randomized Trial," *Journal of the American Medical Association*, 300 (2008), 2631–2637.

Warburton, Darren E. R., Crystal Whitney Nicol, and Shannon S. D. Bredin, "Health Benefits of Physical Activity: The Evidence," *Canadian Medical Association Journal*, 174 (2006), 801–809.

Westfall, Peter H., and S. Stanley Young, *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment* (Hoboken, NJ: John Wiley & Sons), 1993.