# Doing Well by Making Well: The Impact of Corporate Wellness Programs on Employee Productivity[*]

Timothy Gubler
School of Business Administration, University of California, Riverside
*tgubler@ucr.edu*

Ian Larkin
Anderson School of Management, UCLA
*ian.larkin@anderson.ucla.edu*

Lamar Pierce
Olin School of Business, Washington University in St. Louis
*pierce@wustl.edu*

**Forthcoming in Management Science**

June 28, 2017

Abstract: This paper investigates the impact of a corporate wellness program on worker productivity using a panel of objective health and productivity data from 111 workers in five laundry plants. Although almost 90% of companies use wellness programs, existing research has focused on cost savings from insurance and absenteeism. We find productivity improvements based both on program participation and post-program health changes. Sick and healthy individuals who improved their health increased productivity by about 10%, with surveys indicating sources in improved diet and exercise. Although the small worker sample limits both estimate precision and our ability to isolate mechanisms behind this increase, we argue that our results are consistent with improved worker motivation and capability. The study suggests that firms can increase operational productivity through socially responsible health policies that improve both workers' wellness and economic value, and provides a template for future large-scale studies of health and productivity.

**Keywords: Worker Productivity, Health, Wellness Program, Presenteeism, Corporate Social Responsibility**

[*] We thank Hengchen Dai, Andrew Knight, numerous conference participants, and the editorial and review team for their thoughtful comments and recommendations. Doctors Aaron Miller, Kurt Munzer, and George Toth provided valuable medical advising. Executives and employees at *LaundryCo* and the program vendor generously provided time and data for the paper. This research was supported by the Center for Research in Economics & Strategy (CRES) at the Olin Business School. All mistakes are ours alone. Please send all correspondence to ian.larkin@anderson.ucla.edu.

# 1. Introduction

Companies increasingly invest in employee health and well-being (Ton 2014). A recent survey found that around 90% of companies use corporate wellness programs that can include simple biometric screenings such as basic blood tests, advanced screening for diseases such as cancer, exercise programs, nutritional and diet programs, health history and habits surveys, and training on protecting and improving health (Medical Billing and Coding 2012). The prevalence of such programs is unsurprising given growing rates of obesity, diabetes, and other health problems, and the implications of these issues for employer-sponsored health insurance and absenteeism (Baicker et al. 2010; Boles et al. 2004). Obesity has steadily increased to almost 35% of the United States population in 2012 (NCHS 2012), diabetes cases have more than quadrupled since 1980 (CDC 2014), and exercise rates and eating habits have not improved (Gallup-Healthways 2012). This decrease in employees' physical health is reflected in the 131% increase in health insurance premiums from 1999-2009, the cost of which is largely borne by employers (Kaiser/HRET 2012).

Extensive research in the fields of medicine, public health, and health economics shows that the costs of corporate wellness programs are dwarfed by reductions in insurance costs and absenteeism. A recent meta-analysis found that each dollar spent on wellness programs saves \$3.27 in health care costs and \$2.73 in absenteeism costs (Baicker et al. 2010). Although these gains are substantial, they ignore an important class of operational benefits from investing in employee health and well-being—worker productivity. As management and operations scholars argue (Danna and Griffin 1999; Goldstein 2003; Grant et al. 2007; Neumann and Dul 2010; Ødegaard and Roos 2014), healthier employees are not only less expensive and less absent, but are also more productive. A large occupational health literature on "presenteeism" – working while ill – suggests lost productivity costs firms over \$150 billion, almost three times absenteeism costs (Stewart et al. 2003). Despite this value, existing research has not causally linked objective health data from wellness programs with measurable worker productivity changes. This may explain firm confusion about how much wellness programs actually help their bottom line. A recent article in the *Wall Street Journal* notes that "employers are stymied by the difficulties of measuring the financial and health impact of wellness programs" (Weber 2014).

In this paper, we provide the first causal evidence based on objective data that wellness programs and their related health improvements can increase worker productivity. We use novel longitudinal data on individual health and productivity from an industrial laundry company (which we call *LaundryCo). LaundryCo* provided free annual biometric screenings to all full-time employees in four laundry plants; a fifth plant did not participate because it offered a different insurance plan. The voluntary program increased awareness of and attention to health by providing each employee with a personalized health packet detailing current health status and providing recommendations for improvement. Detailed daily production data, combined with annual medical data including bloodwork, lifestyle choices, and vital statistics, allow us to exploit this quasi-experiment to generate causal estimates of the wellness program's productivity impact. The program appears to have improved average worker productivity by over 4%, approximately equal to adding an additional day of productive work per month for each employee, although this estimate is imprecise. We find more precise evidence of productivity gains among those who improve their health that suggest increased work capability. Sick employees whose health improved showed a 10% productivity increase. Strikingly, already healthy employees who improved their health showed an 11% productivity gain. Survey and blood results indicate that these health improvements stemmed from physical activity increases, attention to diet, and other lifestyle changes. In addition, employees with no health problems and no health improvement showed a 6% productivity increase following the program's introduction. We conservatively estimate a return on investment of 76% purely based on productivity gains, with a much higher return if the company were able to reduce non-participation in the program and reduce turnover.

Our paper contributes to several key literatures in operations, management, and strategy. First, we add to growing research in management on the relationship between employee well-being and organizational performance, providing the first causal evidence linking a multi-year panel of medical data to actual individual productivity improvements in a firm. This supports existing cross-sectional and self-reported data in prior work (e.g., Burton et al. 1999; Goetzel et al. 2003; Ødengaard and Roos 2014).

Second, our paper joins the growing literature in operations that uses detailed micro-level production and service data to study environmental, social, and psychological factors that impact worker productivity. Recent studies demonstrate that operational policies such as scheduling and staffing (Chan et al. 2014a; 2014b; Dai et al. 2015; Huckman and Pisano 2006; Huckman and Staats 2011), monitoring and transparency (Bernstein 2012; Buell et al. 2017; Pierce et al. 2014; Staats et al. 2016; Tan and Netessine 2015), performance recognition (Gubler et al. 2016; Larkin 2011; Song et al. 2017), and workflow (KC 2013; Kuntz et al. 2014; Staats and Gino 2012; Tan and Netessine 2014) can all affect worker performance. Even environmental factors outside managerial control such as weather (Lee et al. 2014) may determine productivity. Our study is unique in this literature not only by linking health and productivity, but also by supporting the argument that firm policy can improve operations through worker health and well-being (Ton 2014).

Third, our paper contributes to the large and often conflicting literature on corporate social responsibility (CSR) and firm performance (Margolis and Walsh 2003; Orlitzky et al. 2003; Kitzmueller and Shimshack 2012) by detailing mechanisms through which firms can "do well by doing good." The empirical challenges in this research are widely recognized (Barnett and Salomon 2006; Chatterji et al. 2009), prompting calls to investigate the micro-foundations of CSR (Aguilera et al. 2007). Our study responds to this call, showing how one frequently-used CSR policy can improve worker productivity and firm outcomes.

Finally, our paper contributes to a vast, but largely correlational, medical and public health literature by linking objective health improvement data with objective worker productivity gains in a quasi-experimental setting. This distinction is important for two reasons. First, people are often dishonest or biased about their own productivity (or that of others) (Podsakoff et al. 2003). Second, self-reported productivity is difficult to quantify or value, particularly when measured with scales.

Although our combination of objective health and productivity data provides unique evidence linking employee health initiatives to productivity, we caution that our sample of 111 workers limits our statistical power. Many of our parameter estimates are only weakly significant at traditional levels, and we emphasize that this paper should be viewed as preliminary evidence of how wellness programs improve both health and productivity. We hope that this evidence will motivate companies with larger workforces to collaborate in research that might more precisely estimate the effects observed here.

## 2. The Potential Impact of Wellness Programs on Productivity

### 2.1 Evidence Linking Socially Responsible Employee Policies to Performance

A large literature suggests that socially responsible (CSR) employee policies, such as wellness programs, might improve firm performance, focusing on predictors of firm-level CSR, the outcomes of CSR, and mediators and moderators of the CSR-outcomes relationship (Aquinis and Glavas 2012; Wang and Qian 2011). These studies correlate CSR with outcomes that include improved reputation, consumer loyalty, and attractiveness to investors. Despite this research, however, scholars continue to disagree about the importance of such effects because few papers provide *causal* tests of CSR and firm performance (Flammer 2015).

One stream studies how employee-focused CSR policies might increase performance, finding correlations with employee behaviors and attitudes that include organizational identification (Carmeli et al. 2007), employee engagement and citizenship (Glavas and Piderit 2009; Jones 2010), employee relations (Agle et al. 1999), and attractiveness to job seekers (Turban and Greening 1997). Field studies almost exclusively lack strong causal evidence, with only a few exceptions—none of which study worker health. Burbano's (2017a; 2017b) online labor market experiments show that socially responsible messages increase effort. Carnahan et al. (2017) uses the September 11 terrorist attack to link pro bono work to law firm retention. Flammer and Luo (2017) links retention and engagement to CSR using shocks to unemployment insurance benefits.

This paper focuses on employee wellness programs—one of the most common socially-responsible policies targeting employees. One reason these programs are so common is that the immediate, salient, and measurable benefit from reduced insurance premiums provides managers with easy financial justification for program introduction in the face of internal political opposition (Berry et al. 2010). Additionally, firms may see immediate tangible gains from healthier employees through reduced absenteeism, injury, and worker compensation claims (Chapman 2012). Widespread evidence shows that decreasing a health risk such as smoking can significantly reduce insurance and absenteeism costs (Burton et al. 1999; Goetzel et al. 2003).

Existing research therefore has primarily studied firm wellness program benefits via reduced costs from insurance, absenteeism, and risk, rather than from worker productivity. From an empirical perspective, this focus is understandable. The link between wellness programs and productivity is difficult to causally measure. Matched objective productivity and health data are difficult to obtain from firms, and isolating the treatment effect of such programs amidst other policy changes can be daunting. Furthermore, most companies offer wellness programs to all employees, which means researchers cannot untangle temporal productivity changes caused by time trends or shocks affecting all workers. These factors may explain why meta-analyses of financial returns from wellness programs include almost no productivity-based returns (Chapman 2012).

### 2.2 Mechanisms

We argue that a corporate wellness program could affect employee productivity through two classes of mechanisms: job motivation and capability. In the next sections, we explain that the likely importance of motivation and capability in productivity gains depends on two observable employee characteristics—pre-existing health problems and health improvement during the program. Although this study cannot identify the sources of employee health problems or improvement, these individual characteristics can help illuminate whether motivation or capability improvements likely drive our observed productivity gains.

Figure 1 shows four categories of employees likely to improve productivity through these mechanisms. Participating employees can be categorized based on two observable characteristics across time: if the program identified a health problem and whether the employee improved their health between exams. Because even healthy employees can improve their health, this classification generates four participant types. We argue that all four types might increase productivity because of motivation based on job satisfaction. Those with health problems might enjoy further motivational gains due to reciprocity based on feelings of gratitude. Additionally, workers who improve their health might increase productivity because of improved physical and mental capabilities. Previously unhealthy employees who improve their health might see the largest capability gains because of existing impediments of stress, pain, and physical weakness.

<<< INSERT FIGURE 1 HERE >>>

**2.2.1 Productivity Improvement Through Motivation: Increased Job Satisfaction**. One way that a wellness program might improve motivation and therefore productivity is by demonstrating an employer's concern for employee well-being, which spans many psychological, physical, and social dimensions (Grant et al. 2007). The costly implementation of a program designed to improve well-being might improve worker attitudes by credibly signaling to employees the firm's broader concern for the quality of their work life, and even the quality of their life outside the workplace. The program in our empirical setting, for example, costs approximately $240 per year for each employee.

Employers frequently offer corporate wellness programs to all employees without knowing who will benefit. Even workers themselves may not be able to predict, *ex ante*, whether they will benefit. Therefore, *all* participants, not just those who receive new information from the program, might perceive increased organizational support and experience increased motivation through job satisfaction. Indeed, wellness programs, independent of efficacy, have been shown to increase job satisfaction (Zoller 2004) and raise perceptions of organizational commitment towards employees (Parks and Steelman 2008). Both job satisfaction and perceived organizational support have been positively associated with job performance (Armeli et al. 1998; Yee et al. 2008).

**2.2.2 Productivity Improvement Through Motivation: Gratitude and Reciprocity**. While all employees might improve motivation via job satisfaction, employees who learn they are ill might improve motivation more than their healthy colleagues. One of the major aims of a wellness program is to help employees identify and focus attention on existing conditions and illnesses. Almost half of overweight

Americans don't recognize their weight problem (Ingraham 2016), while one in four Americans with diabetes are unaware of their condition (CDC 2014). All such employees, regardless of whether the program also helps them remediate their illness, might feel gratitude to the employer for providing information about existing but unknown health conditions. Since this information is inherently valuable, employees who receive this gift will be inclined to reciprocate (Bartlett and DeSteno 2006; Tsang 2006). Reciprocity theory holds that actors such as employees react to unexpected giving by responding in turn, even if a receiver does not want or will not use a valuable gift (Grant and Gino 2010; Adams et al. 2012). When an employee receives a work benefit, she may strive to relieve this imbalance though contributions to the organization (Eisenberger et al. 2001). One natural way employees could reciprocate is via increased productivity. Although all participants might feel some gratitude, the value of the information provided by the program is highest for workers with preexisting health problems, making these workers most likely to feel gratitude and thus reciprocate (Dabos and Rousseau 2004; Hekman et al. 2009).

**2.2.3 Productivity Improvement Through Work Capability.** A wellness program may also increase productivity by helping employees improve their health, thereby strengthening their work capability. As noted above, employees might learn of an existing health condition, realizing the need to improve their health. Yet even if they were aware of a health problem, the program may focus their attention on health in ways that nudge them toward action. Free counseling on nutrition, substance abuse, weight, exercise, and other health habits provided through the program may help all employees to make positive lifestyle changes through several behavioral mechanisms. First, counseling from the program may produce the type of concrete improvement plan known to improve health behavior (Milkman et al. 2011; Dai et al. 2012; Rogers et al. 2015; Beshears et al. 2016). Second, the program may provide an effective reminder that establishes better habits (Calzolari and Nardotto 2017). Recent research shows that subtle nudges about lifestyle choices and health and longevity can indeed increase commitment to healthy choices in diet, exercise, and sleep (Vallgårda 2012). The specific actions that employees take in a wellness program – seeing a registered nurse, getting blood drawn, and receiving information on typical health problems – may provide one such nudge.

Employees who improve health problems would likely see the largest productivity gains because of substantially improved physical capability from remediating health issues. There is consensus in the occupational health literature that poor health reduces work capacity and substantively changes wages, hours worked, labor force participation, job choice, turnover, retirement, and occupational choice (Currie and Madrian 1999). There is scant evidence, however, of the direct link between health improvement and on-the-job productivity at the individual level. Instead, researchers have indirectly connected health to productivity either through human capital development identified by educational attainment (Becker 2007; Conti et al. 2010) or through national and other macro-level measures (Currie and Madrian 1999). But even healthy employees without diagnosed problems might improve health through adopting healthy lifestyle choices in areas such as diet, exercise, and sleep. Each of these has been linked to employee on-the-job productivity through capabilities based in stamina, energy, and mood (Thayer et al. 1994). A substantial literature on presenteeism—present workers impeded by illness, depression, injury, or pain—links self-reported productivity loss to mental and physical health conditions that include diabetes, depression, anxiety, cancer, migraines, and arthritis (Stewart et al. 2003). For example, surveys from Lockheed Martin correlated productivity losses with health problems such as migraines (4.9% loss), allergies (4.1% loss), asthma (5.2% loss), influenza (4.7% loss), and depression (7.6% loss) (Hemp 2004). Even when health improvements are not specifically tied to the capability to carry out a critical job task, they may increase worker task productivity due to improved mental health and reduced distraction from pain and discomfort. Recent evidence from Christian et al. (2015) suggests that pain reduction, and subsequent energy improvements, strongly affects discretionary tasks such as productivity or prosociality.

Thus, while we expect that general health improvements through lifestyle choices could increase productivity through better job capability for all workers, we would expect these gains to be largest for those with existing health problems.

## 3. Data and Methodology

The major innovation in our data is the ability to link objective employee health data with daily individual productivity measures in a quasi-experimental setting. Previous wellness program studies have either lacked a control group or used purely cross-sectional or survey data. Our paper features both a quasi-control group and longitudinal objective data for both productivity and health paired with self-reported survey and demographic data for 111 workers in four treatment plants and one quasi-control plant. The major limitation of our data, as we will detail in Section 4.2.1, is our relatively small sample, which limits our statistical power to precisely estimate parameters.

### 3.1 Setting

Our data originate from a private industrial laundry service company that we call *LaundryCo*. *LaundryCo* is one of the largest independent industrial launderers in the United States, providing uniforms, mats, and other garments to clients that include auto shops, construction companies, restaurants, and hospitals. Items are regularly laundered and repaired by *LaundryCo* using an innovative IT and production system ensuring timely delivery. The cleaning and repair of products is carried out primarily by workers in five plants across four states. These plants are nearly identical in layout, equipment, staffing positions, and products, and have about 15 production workers employed per laundry line on a regular day. Three of the plants, including our control plant, have two lines with about 30 workers per day, and the other two plants have a single line with about 15 workers per day. Plants also employ other workers for whom we cannot measure productivity, including managers, delivery workers, and support staff. The average production worker stays approximately four and a half years, with a median tenure of two years. The primary difference between our plants is the unionized workforce at the control plant, which reduces turnover. This difference is fixed throughout the period of our study, and therefore will not bias our identification of the wellness program's treatment effect.

Soiled clothes, mats, and linens are dropped off at the plants and proceed through a complex sequence of sorting, cleaning, drying, pressing, and repair before being loaded and returned to customers. Non-uniform items such as floor mats or linens go through their own unique process. Workers are cross-trained on multiple tasks, but typically specialize in only a few. Similar to other service operations, worker efficiency is crucial to *LaundryCo* profitability. Mistakes and bottlenecks create costly production line disruption and leave downstream workers with insufficient workload. Mistakes may include incorrect initial sorting, insufficient cleaning, or failure to repair garments before the final quality check. Bottlenecks result both from mistakes and the inefficiency of upstream workers. For instance, if the dryer operator falls behind, the pressing machine operator might be idle until garments arrive for pressing. Downstream worker dependence on productivity in upstream colleagues magnifies presenteeism costs for *LaundryCo*.

In Spring 2010, *LaundryCo* hired an outside company ("the vendor") to provide a free wellness program to employees. Management's goal was to reduce insurance premiums and improve employee health. The program was offered to all employees as a voluntary free benefit, with management actively encouraging participation. Excluding managers and non-factory workers such as salespeople, 136 employees chose to participate, while 31 chose not to. *LaundryCo*'s human resource director told us that employees chose not to participate because of either a spouse's insurance coverage, absence on the day of the program, a fear of doctors, or worry that the program would uncover drug use (although the program did not test for drug use). Participating employees received a 15% monthly insurance premium decrease (about $1.75 to $11 per month), which represented about 0.4% of monthly wages for the average worker.

We note that the carefully-designed *LaundryCo* program adheres to two design principles that we will address in the discussion section. First, participation was purely voluntary. While participants received a small insurance premium reduction, non-participants' rates did not change. *LaundryCo* management did not record non-participants, and plant managers report not knowing the identity of non-participants. Second, *LaundryCo* did not have access to the health data. We received these data directly from the vendor, and *LaundryCo* was not party to any individual results from the program. All employees were informed that participation was voluntary and that privacy was assured because *LaundryCo* would not administer the program.

The program began with a blood draw at the plants, which defines our treatment date. The blood samples were then shipped to LabCorp, a national medical laboratory testing company, and tested for 42 common health markers. Employees were also given a health survey from Wellsource, a provider of evidence-based health assessments. The survey asked about health background, nutrition, fitness, stress level and mood, drug use, and other behaviors. Because it was administered during work hours, and reviewed by the nurse, the survey had few missing answers (2.2% of responses) that necessitated data exclusion. Approximately three weeks after the screenings, nurses from the vendor held an educational seminar at each plant and presented each worker with a personalized report detailing their results. About 97% of participants in our sample had at least one abnormal test result each year, although some of these reflected small, likely-random variations on a few blood test measures. Both *LaundryCo*'s HR director and interviews with workers indicated that because many workers rarely visit doctors, many biometric screenings uncovered serious or unknown health problems.[1] The nurse called approximately 20% of employees with severely abnormal results within a week of the first visit, explained the results, then asked for permission to forward the results directly to the employee's physician. Employees who lacked a primary care physician were offered a referral. The health packet detailed the individual's health status, including blood results and abnormalities, health behavior scores, anticipated future risks, and personalized suggestions for improvement. Even employees who were aware of an existing medical condition, such as obesity or borderline diabetes, might not have known its deleterious impact, or ways to remediate it, until they received this detailed, individualized information.

*LaundryCo* offered the full program, including biometric screenings, surveys, and counseling, to all employees in 2010, 2011, and 2012. Although the program continues to operate for new hires, *LaundryCo* stopped offering the full biometric screenings for previous participants starting in 2012.

### 3.2 Data

The data span 2009-2012 and cover all production workers at five *LaundryCo* plants for which productivity data are available. Figure 2 illustrates the program structure with three sets of visits from the outside vendor. *LaundryCo* offered the wellness program to all employees in 2010, 2011 and 2012. To measure whether an employee's health improves due to the wellness program, we limit our sample to employees who participated in two screenings in a row, and limit our analysis to the first such instance (i.e., 2010 and 2011, or 2011 and 2012). The date of treatment is defined as the date of blood draw for the first instance of participation. Turnover severely reduces our sample. Of the 347 production workers employed during our sample period, 56 stayed for less than a month, while an additional 106 stayed for only one instance of the program (98 in the treatment plants and 8 in the control plant).

Our dataset combines three main sources. First, we received health data from the outside vendor. These data include objective biometric blood tests and wellness survey results. Second, we employed outside physicians to use the biometric and survey data to assess each employee's health level and improvement between years. Finally, *LaundryCo* provided demographic, human resource, and daily productivity data from their IT system. We merged and de-identified these three datasets under an approved university IRB protocol.

<<< INSERT FIGURE 2 HERE >>>

**3.2.1 Biometric data.** The vendor compiled longitudinal blood data for each participant, including a panel of 42 blood tests. These biometric screenings assessed abnormalities in diabetes, cholesterol, kidneys, enzymes, iron, electrolytes, cell balance, thyroid, complete blood counts, white blood counts, and prostate (PSA). The average worker had nearly five abnormal blood results out of 42 tests given. However, since the

[1] There were many employees who explained how the wellness program helped improve awareness of their health. One employee said, "They caught my thyroid! They let me know how out of balance I was and made referrals to help get me on the right medication." Another stated that, "Our wellness program helped me find out that I had diabetes…. Since the test I have been treated and am feeling much better overall." A third employee shared that, "During the blood draws last year it was discovered that I had high potassium levels and also had a hyperactive thyroid…. Since then I have got with my physician and I am feeling much better." Finally, an employee shared this story: "[The wellness program] opened my eyes to something I did not know I had. I was unaware I was diabetic or had high cholesterol. Since last year, I went to the doctor and have been put on medication and my diabetes and cholesterol is [sic] under control. I have also lost 25 lbs. If this had gone unnoticed, I could be going down the same road as my brother who has recently had to have toes removed due to his diabetes."

normal test ranges represent 95% confidence intervals, and since the tests themselves contain measurement error, even healthy employees were likely to have at least a few abnormal results out of 42 tests. In addition, the outside vendor measured blood pressure and calculated body mass index (BMI), a measure of obesity.

**3.2.2 Survey results.** The vendor also provided survey results from the Wellsource wellness questionnaire, which participants completed at the time of testing and included 110 questions on health history and behavior such as exercise and eating habits, drug use, sleep behaviors, current medications, mental health, job satisfaction, health learning interests, and safety (e.g., seatbelt and sunscreen use). Figure A1 in the Appendix shows the actual survey. The survey elements used in our analyses had only 2.2% of responses missing, resulting in data exclusion.

**3.2.3 Physician evaluation of biometric and survey data.** We employed three experienced internal medicine residents from a major Midwestern university hospital to thoroughly evaluate the pre-existing health and health improvements of each employee. The doctors evaluated health and improvement first individually and then as a group. We hired physicians because, while we have detailed health information on each employee, there was no overall assessment of illness, or whether overall health improved between testing events. Physician focus groups told us that it would be difficult to build objective measures of sickness based on the data alone, and that doctors are trained to examine a mix of objective measures and subjective evidence such as responses to questions during office visits. Based on this feedback, we chose to have a group of physicians read the entirety of each participating employee's file.

We asked each physician to individually answer four questions for each patient by evaluating their complete biometric and survey records (see Figure A2 in the Appendix for questionnaire): (Q1) Do they likely have one or more medical conditions? (Q2) How seriously ill is the patient (5-point scale)? (Q3) How much would that problem impact their ability to carry out eight hours of manual labor (5-point scale)? (Q4) How much did the employee's health improve between annual tests (5-point scale)?

Inter-rater agreement (IRA) on Q1 was good, with Fleiss' Kappa of 0.50. This means that for 90% of the employees, all three doctors marked them as having a medical condition, or not having one. Inter-rater reliability (IRR) levels on Q2-Q4, however, were much lower, with Krippendorff's Alpha scores of 0.17, 0.12, and 0.20. These values would be low for non-expert raters on simple tasks, but are consistent with many studies of expert agreement and IRR in the medical literature (e.g., Einhorn 1972; Elmore et al. 1994; Beam et al. 1996; Chiang et al. 2007). For example, a study assessing the diagnostic abilities of 32 medical interns by 12 full-time medical faculty using 128 different clinical evaluation exercise methods found IRRs ranging from 0.00 to 0.63, with a mean of only 0.23 (Kroboth et al. 1992). This means that medical faculty observing the exact same interaction between an intern and a patient were in agreement about how successfully the intern diagnosed a patient problem at close to the same rate as our physicians agreed on Q2-Q4, and at a much lower rate than our physicians agreed on Q1. With results from so many blood tests as well as survey results, nearly every patient-year observation contains some contradictory data, and physicians tend to focus on different measures based on their own beliefs and experiences (Eddy and Clanton 1982). Furthermore, most studies of IRR in medicine involve a physical examination of the patient, which our physicians lacked. Therefore, our low IRR ratings are consistent with the literature on the difficulty in medical evaluations.

In fact, the physicians themselves expressed to us that Q2, Q3 and Q4 were difficult given the large amount of data and many borderline results, and after completing their individual ratings, they proposed collective evaluations involving a scoring system established by the doctors, with higher values broadly reflecting worse health (see Figure A3 in the Appendix for the scoring system). We use both the individual and collective measures for Q4, and the results are highly similar. We chose not to use Q2 in the study for several reasons. First, Q2 is simply a more detailed version of Q1, yet this detail produces poor agreement across physicians and is therefore inferior to Q1. Second, the physicians indicated frustration with this question and that the dichotomous choice of Q1 was more appropriate for general health assessment. We excluded Q3 because the low IRR likely resulted from forcing physicians to match general health assessments to job tasks they had not observed. In the case of both variables, the low IRRs indicate significant measurement noise that would be prohibitive in a study that includes relatively few employees.

**3.2.4 *LaundryCo* IT data.** *LaundryCo* provided data at the employee level on individual productivity and demographics for 2009-2012. Productivity data are at the worker/day level, and the other data are measured as of January, 2013. Demographic data include employee age, salary, tenure, and plant assignment—all of which are absorbed in our fixed effect regressions. Worker productivity is based on how long an employee works each day and how efficient they are on each given task.

### 3.3 Identification Strategy

We use a difference-in-differences (DiD) study design. The DiD design treats employees in *LaundryCo*'s four plants that participated in the wellness program as the treatment group and the single non-participating plant as a quasi-control group (hereafter called the control group). Because some treatment plant employees chose not to participate in the program, we estimated a separate treatment effect for these employees. The DiD strategy "differences out" fixed differences between treatment and control groups, and uses post-treatment changes for the control group as a counterfactual for what would have happened had treatment group individuals not participated in the wellness plan. The DiD approach is the most widely used methodology to examine the impact of exogenous shocks or policy changes (Gertler et al. 2011).

Our treatment group is comprised of 69 employees who were present for at least two wellness program exams at one of the four participating plants. We exclude from this group 81 employees who left employment with *LaundryCo* after participating only once in the wellness program because our identification relies on observing an employee's health improvement in a second exam after their initial participation a year prior. We furthermore drop 17 treatment group employees who opted out of participating in the wellness program a single time, and who left the company before a second program could be administered. We also drop 67 treatment group employees who were not employed during a single evaluation. Our final treatment group consists of two types of employees: the 55 workers who chose to participate and the 14 who did not, nine by choice and five because of absence on exam day.[2] Following the literature on health intervention, we term participating workers "compliers," and the non-participants "non-compliers." Both compliers and non-compliers were employed at a treatment plant and received information about the program. Although non-compliers were younger and had shorter work tenure, their average efficiency, pay, and daily hours worked were not substantially different from compliers, as shown in Appendix Table A1. The control group is comprised of workers from *LaundryCo*'s unionized fifth plant, which used a different insurance plan and did not participate in the program. As with treatment group workers, we limit the control group to the 42 employees who were employed for the period encompassing at least two exams at treatment plants.

Although DiD strategies do not require identical treatment and control groups, in our case workers from both groups were similar on most dimensions. All five plants work on the same tasks, use the same production technology, and share common floor layouts. Three of the plants – the control plant and two treatment plants – have approximately 30 employees working on a given day. The other two treatment plants average 15 workers, but simply have fewer lines within a given category (e.g., one mat rolling station instead of two). Also, two of the treatment plants are geographically proximate (32 and 34 miles) to the control plant, which addresses local shocks such as weather that might affect productivity. Table A1 in the Appendix presents pre-program demographics and efficiency for the three groups. As we note earlier, the one key difference at the control plant is its unionized workforce and consequent lower turnover. Since our DiD with worker fixed effects will absorb this fixed difference, this does not invalidate our identification strategy. If the program itself led to changes in turnover our results would be biased; however, as seen in Appendix Figure A4, there is no discernable impact on turnover.

Interviews with both corporate and control plant managers confirmed that no wellness program was offered there. Interviews also revealed that nearly all other major plant policies were directed by the corporate office, and affected all plants, not just the control or treatment plants. There was one exception – an attendance awards program was put in place in the second year of our study at the control plant. We discuss the implications of this program in section 4.2.3. Of course, managers may have their own management style, but any fixed differences in management policies (or any other variable) do not affect DiD estimates.

DiD studies do not require random assignment of treatment and control, but a potential weakness of our study is the lack of a true control group, as employees choose the plant at which to seek employment. However, it is extremely unlikely that participants chose their plant based on expectations of a future wellness program. Interviews suggest employees desire to work at the closest plant, and very few employees would find it desirable to transfer to a different plant from their current location. Therefore, with respect to being offered the wellness program opportunity, our intervention approaches random assignment.

### 3.4 Variables

[2] Since the exams were known ahead of time by workers, employees may have chosen to be absent to avoid the exam, a possibility acknowledged by management.

**3.4.1 Dependent variable.** Our dependent variable is daily worker efficiency. *LaundryCo* uses a sophisticated IT system that carefully tracks each worker's productivity (called "efficiency") on each task every day. To do so, it measures the garment processing rate for each worker compared to the time-studied expected rate, as determined by corporate headquarters. Scores are normalized such that 100 reflects performance that meets expectations. For example, the time-studied rate for pressing dress shirts is 50.4 seconds, meaning an employee must press over 71 shirts each hour to earn a score of 100. The system computes an overall daily efficiency rate for each worker, equal to the weighted average (by time spent) of the worker's efficiency scores on each task that day. For example, a worker who spent two hours sorting soiled clothes with an efficiency score of 80, two hours loading soiled clothes and the appropriate soap into washing machines with a score of 140, and four hours unloading clean but wet clothes into bins to be taken to the dryer area with a score of 160 would have a final daily efficiency of 135. A typical employee will have an efficiency number around 110-120, with high performers consistently performing above 130 and low-performers at less than 100.

**3.4.2 Independent Variables.** Our independent variables represent four broad constructs: *Compliers*, *Non-compliers*, *Sick*, and *Better*. *Compliers* takes a value of 1 for treatment plant employees who participated in the program twice, and a 0 for non-complying and control group employees. *Non-compliers* takes a value of 1 for employees who chose not to participate in the program when offered, or were absent, and a value of 0 for compliers and control group employees. Empirically, these two variables are interacted with a dummy for the post-treatment period, which for each employee is the date of initial wellness program participation.

In our primary health specifications, *Sick* is a dummy variable created from question 1 in the physician health evaluation. Because IRR on this question is high, indicating 90% unanimity among the three doctors, we require all three physicians to designate a worker as having a health condition based on the results of the wellness intervention that year. Results were highly similar using a cutoff of two of three physicians designating a worker as sick. There were 36 "sick" employees, and 19 "not sick" employees.

Similarly, in our main specifications, *Better* is a dummy variable using question 4 from the physician evaluation that indicates significant health improvement between the first and second exam. We designate an individual as *Better* if the average doctor score indicates health improvement between periods (i.e., average >3). Fifteen workers qualified as "better," and 40 workers were "not better." Note that both "sick" and "not sick" employees were classified as "better." The largest of the four worker categories in Figure 1 was the 27 "sick, not better" employees. Thirteen workers were "not sick, not better," six were "not sick, better" and 9 were "sick, better." As stated previously, there were 14 total non-compliers at the treatment plant employed during at least two blood draws.

We examined the use of continuous rather than binary measures of both *Sick* and *Better*, and while overall model fit was superior, little additional insight is derived. We therefore present the results using the binary measures, which are much easier to interpret. We present results using continuous measures in Appendix.

**3.4.3. Control Variables.** One of the main benefits of the DiD methodology is that fixed differences in the control and treatment units do not affect the treatment estimate. Because fixed, pre-existing differences between groups are "differenced out" of the treatment estimate, including them in the regression (for example as a time invariant dummy variable) will not influence the results. Additionally, as is standard when using individual employee data, we include individual fixed effects, such that our treatment estimate represents the change *within* an individual employee. Collectively, these factors mean that we do not require an extensive set of control variables. Fixed differences across employees such as gender, ethnicity, and even average capability are fully absorbed by individual fixed effects and cannot enter our specification.

However, employees do differ over time in their experience level; a given employee may be more efficient several months after starting than they are in their first week. To control for within-employee learning, we control for the cumulative months of job experience. The results are robust to using a dummy variable for employees with less than a year of tenure, which interviews suggest is well after any learning-by-doing ceases.

Our last control is a plant-specific time trend; we use month fixed effects interacted with plant ID.

### 3.5 Specification

Our DID specification model is the following:

$$Y_{ijt} = \alpha_i + \beta_1 * Post_t + \beta_2 * Post_{it} * Non\text{-}Complier_i + \beta_3 * Post_{it} * Complier_i + \beta_4 * Post_{it} * Complier_i * Better_i + \beta_5 * Post_{it} * Complier_i * Sick_i + \beta_6 * Post_{it} * Complier_i * Sick_i * Better_i + \lambda_{it} + \gamma_{jt} + \varepsilon_{ijt}$$

where $Y_{ijt}$ is efficiency for worker $i$ in plant $j$, at time $t$, $Post_i$ is a dummy indicating all days after the first screening for employee $i$; and $Post_{it}*Non\text{-}Complier_i$ and $Post_{it}*Complier_i$ are treatment group employees in the post period who choose not to participate and to participate in the program, respectively. $\lambda_{it}$ is the employee's cumulative number of months of experience at *LaundryCo*, and $\gamma_{jt}$ is a plant specific time trend.

Our specification allows us to test for heterogeneous treatment effects depending on whether the program identifies a worker as sick, and whether an employee's health improves in the year after implementation. The interactions between the *Post* variable (defined as 1 starting the date of the blood draw) and the *Non-Complier* and *Complier* variables show the effect of the program on non-compliers and compliers, respectively, who are neither sick nor better after the program. The subsequent triple and quadruple interaction coefficients represent the incremental effect of the program on compliers who get better ($\beta_4$), are sick ($\beta_5$), and are both sick and better ($\beta_6$). We note that the time-invariant baseline effects, such as $Non\text{-}Complier_i$, $Complier_i$, $Complier_i*Sick_i$, and $Complier_i*Better_i$ are all absorbed by employee fixed effects. We also note that we can collapse complier categories and look at fewer complier groups by not differentiating by *Sick*, *Better*, or both.

All of the treatment coefficients ($\beta_2$ to $\beta_6$) are compared to the control group of workers not participating in the program. Of course, we do not observe whether control group employees are sick or improve their health. However, this does not bias our econometric results, because our *sick* and *better* variables refer to sickness highlighted in the wellness program, and health improvements achieved due to the same program. Control group workers by definition do not learn about their health or make health improvements due to a wellness program, since they did not hear about or participate in any such program. Any changes to their health awareness, or their actual health, serve as the baseline counterfactual against which our wellness program effects are measured. For example, general societal trends towards healthier eating and exercise, or the effects of public policies such as the Affordable Care Act on health, are reflected in the control group and are "differenced out" of the treatment estimate. The control group models the counterfactual of what would have happened to treatment group employees absent the intervention.

We use ordinary least squares (OLS) to estimate our DiD model, clustering standard errors at the individual level.

## 4. Results

As noted above, 65% of treatment group compliers (36 of 55) were classified as sick by the physicians, and 25% of sick workers improved their health after the intervention (9 of 36). Examples of common health abnormalities were high cholesterol, obesity, hypertension, chronic pain, and self-admitted drug abuse. While a 65% rate of sickness may seem high, the Center for Disease Control in the United States estimates that 50% of adult Americans have at least one chronic health condition (Ward et al. 2014), although they note this is a conservative estimate and does not include mental health, drug use, or obesity, which our doctors included. Because of this broader definition of sickness, and because our sample is almost exclusively low-income workers who have higher rates of illness than the general population, this higher sickness rate is not surprising. Table 1 presents descriptive statistics for our main variables of interest, and Table 2 provides correlations for the final sample. Note that Table 1 is at the worker-day level, which matches our regressions; the statistics above on overall rates of sickness are at the worker level.

<<< INSERT TABLES 1 and 2 HERE >>>

Table 3 shows the formal statistical results from our regressions. Model (1) shows the overall effect of the program on compliers and non-compliers; model (2) breaks down compliers into sick and not sick employees; model (3) breaks down compliers into better and not better employees; and model (4) shows the model with all four employee types. These models all use the independent doctor assessments to define the *Sick* and *Better* variables; to be defined as *Sick,* all three doctors must judge a given employee to be sick based on data from the initial blood draw. Appendix Table A2 shows the same models using the doctors' collective scale to define the *Sick* and *Better* variables, and the results are highly similar.

Rather than present the raw regression output, which shows the marginal effect of being in each subsequent employee category (e.g., the marginal effect of being "better" on top of being "sick" and a "complier"), Table 3 shows the total effect for each employee type, which represents the linear combination of all relevant interaction coefficients. These should be interpreted as the average post-treatment productivity change for each group relative to the control group. Marginal effects of model (4) are in Appendix Table A3.

The point estimate on model (1) suggests that the wellness program improved productivity among compliers by 4.89 points, which is a 3.9% increase from an average productivity level of 125 points. Non-compliers had reduced productivity of -7.21 points or 5.8%. Both estimates are statistically insignificant at conventional levels (p=0.229; p=0.179). However, models (2) and (3) both indicate that some employee types saw statistically significant increases in productivity. In model (2), the productivity of non-sick employees increases by 9.4 points or 7.5% (p=0.054), while model (3) estimates that the productivity of better employees increases by 12.7 points or 10.2% (p=0.036). Results for those that do not improve are small and very imprecise. In both models (2) and (3), the effect of the program on non-compliers is similar to model (1).

Finally, model (4) estimates separate treatment effects for the four groups in Figure 1. Because the number of employees of each type is smaller than in models (2) and (3), the estimates are less precise, but the results are fully consistent with the earlier models. Specifically, non-sick, better compliers ($\beta$=13.461, p=0.073) and non-sick, non-better compliers ($\beta$=7.662, p=0.079) both see improved productivity. Sick, better compliers also see better productivity, although the estimate is even less precise ($\beta$=11.778, p=0.106). Sick, non-better compliers see almost no identifiable change in productivity ($\beta$=-3.068, p=0.468).

<<< INSERT TABLE 3 HERE >>>

The results from model (4) using all four employee groups are consistent with several of the mechanisms that we presented earlier driving productivity improvements. First, improved job satisfaction or commitment may explain why healthy compliers who do not get healthier still increase productivity. Second, even larger productivity gains among both the healthy and sick who get better suggest improved capability through either overall well-being or physical improvement. We cannot claim evidence for these mechanisms, but rather present them as possible explanations for important individual productivity gains.

Our null effect for sick, non-improving workers casts doubt on the role of gratitude and reciprocity, but we note several factors that may counteract gains among employees whose health does not improve. First, the discovery of illness, and the failure to remediate it, may increase stress and depression, which have been widely linked to decreased productivity and safety (Kuntz et al. 2014). Second, as we noted earlier, severe health problems may not be easily addressable in the short-term. Although no employees were diagnosed with a terminal illness, several had severe, uncontrolled diabetes, extremely high cholesterol, and morbid obesity, problems might take more than a single year to rectify. While our results provide no evidence of gratitude and reciprocity, employee interviews suggest these mechanisms were present in some cases.

### 4.1 Testing the Identifying Assumption of Parallel Trends

One of the main confounds of difference-in-differences models is that different pre-trends in the control and treatment groups can generate spurious, significant results. The formal test for such pre-trends is to run a "leads and lags" model that estimates separate "treatment effects" for a set of pre-treatment and post-treatment periods (Autor 2003; Angrist and Pischke 2009). While the basic DiD model treats the entire pre-treatment period as the baseline against which the post-treatment period is compared, the leads and lags model uses a smaller pre-treatment period as the baseline. Other pre- and post-treatment periods are then tested for treatment effects against this smaller baseline.

Formally, this regression model is identical to the main regression specification outlined above but with a series of time period dummies and their interaction with the treatment group. The omitted time period provides the baseline difference against which each time interaction is tested. If pre-trends are driving the effect, pre-treatment period coefficients would initially be large in the pre-treatment period and in the opposite direction of the coefficient in the main model, with a marked trend towards the direction of the model's coefficient that begins well before the time of treatment. The lags and leads models also reveal if treatment effects start immediately and if they persist.

Figure 3 shows quarterly leads and lags results for all five employee types in model (4) of Table 3: the four complier types and non-compliers. The omitted quarter is the three months before program announcement, since this is the closest period not influenced by the program. The first quarter after the omitted quarter contains the program announcement, the blood draw, and the seminar, that occurred in consecutive months.

For the four complier types, there is no evidence of differential pre-trends—none of the pre-treatment coefficients are large and negative in ways that might spuriously drive a positive coefficient in the baseline model. None of the pre-treatment coefficients for these four types are statistically significant; the coefficients are all less than 8 productivity points and almost always less than 4 points. Compared to the quarter just before implementation, productivity differences between treatment and control groups are thus close to zero. Furthermore, the largest pre-treatment coefficients (Figure 3C representing "not sick, better" employees) are positive, which if anything biases the base DiD model towards a negative result, not a positive one.

<<< INSERT FIGURE 3 HERE >>>

For three of the four complier groups – not sick, not better; not sick, better; and sick, better – there is an immediate increase in productivity differences between treatment and control groups after program implementation, consistent with the program itself causing observed productivity improvements. For sick, better employees, the productivity jump appears to be sustained throughout the five quarters after implementation, suggesting that the health improvements brought about by the wellness program brought sustained productivity increases. For not sick, better employees, the results are not as persistent; while the point estimate remains positive throughout the sample, the effects decrease and become less precise starting nine months after program implementation. We note that this cell of employees is the smallest in the sample with only six employees. For not sick, not better employees, we see an immediate jump in productivity, but this effect tapers off over time. This result suggests that productivity gains from the intervention may be short lived if not accompanied by real improvements in health. Finally, for sick, not better employees, there appears to be a short-term negative impact on productivity.

Plots of the raw data by group also suggest a treatment effect on some but not all workers, and that pre-trends were not different across groups. Raw monthly efficiency averages are plotted for better compliers, not better compliers and control employees in Appendix Figure A5. The version of the chart showing the average pre- and post-treatment efficiency for these groups (as a horizontal line) suggests steady averages for not-better compliers and control employees, but a large average increase for "better" employees. This chart is most similar to the DiD regression approach, for which no time trends are assumed, but we also show raw data charts with linear and polynomial spline best fits. All three charts suggest a flat control group and an unchanged "not better" complier group, with a large, discernable positive jump for "better" compliers.

The lags and leads results on non-compliers (Figure 3E) suggest their productivity was already declining before program implementation. Differences between treatment and control group productivity compared to the omitted quarter were fairly large, positive, and close to statistical significance at the five percent level. The negative result on non-compliers in the statistical models is at least partly attributable to an apparent declining trend in productivity of non-compliers that started well before the wellness program was implemented.

### 4.2 Robustness Tests

We conducted several robustness checks to these main findings. First, Appendix Table A4 shows regression results using block-bootstrapped standard errors at the worker level (Cameron and Miller 2015), with similar results. Second, we carried out a series of placebo tests, randomly assigning employees to different groups, and randomly assigning intervention dates. This is to ensure that the serial correlation errors that plague some DiD specifications are not generating spurious results (Bertrand et al. 2004). Figure 4 shows placebo tests on the sick, better category of employees. For 47 of the 50 placebos run, the coefficient is both smaller in size *and* less statistically significant than the estimated coefficient. Since the coefficient is significant at the 10% level, this is approximately what one should expect from the placebo tests. Appendix Figure A6 shows placebo results on the other 3 complier groups, as well as non-compliers. We also ran robustness checks on the non-complier category. In one such test, we excluded the five non-compliers who were absent from the plant on the day of blood draw, as these employees might have not participated for exogenous reasons such as sickness rather than by choice. These results are in Appendix Table A5. The results for non-compliers is considerably larger and more significant when the five absent non-compliers are dropped; however, the basic results for compliers do not change significantly.

Finally, we examined the impact of using continuous measures for our "sick" and "better" variables. Our doctors provided 5-point scales for both variables, and we used the average of these raw scores in place of the binary variables reported previously. However, there was not a lot of variation within these variables, and interpreting the regression results and continuous interactions becomes very complicated, which is why we used binary measures in the paper's main results. In the Appendix, we show regression results using continuous measures for these two variables in Table A6. Appendix Figure A7 shows a graphical representation of the results, which confirms that employees who are the sickest and whose health most improves see the largest productivity gains.

<<< INSERT FIGURE 4 HERE >>>

### 4.3 Specific Health Improvement Mechanisms

We next examine specific employee actions driving the health and productivity improvements. For these tests, we primarily use survey data to identify workers who improve self-reported behaviors on nutrition, exercise, and stress. We use the survey data to redefine our "*Sick*" and "*Better*" employee categories. We classify individuals as *"Sick"* if their score is below the minimum vendor-communicated threshold for that dimension. We classify individuals as "*Better*" in each area if they improve their self-reported survey scores. We additionally use blood data on HDL cholesterol to measure nutrition and exercise, as it is directly linked to diet and exercise, is difficult to improve by medication alone (Wood et al 1988; Mensink et al 2003), and is an objective rather than a self-reported measure. We note, therefore, that both the "sick" and "better" variables were different from the doctor surveys in this set of results. Regression results using these new measures suggest that productivity gains are driven by lifestyle improvements in exercise, nutrition, and stress. Figure 5 shows that those who are "sick" on these dimensions, but who subsequently improve, see positive productivity gains. All of these coefficients except for "nutrition" are significant at the p<.10 level.

<<< INSERT FIGURE 5 HERE >>>

Figure A8 of the Appendix repeats the analysis from Figure 5, with employees grouped by the original "*Sick*" and "*Better*" variables defined by the three physicians. Although the number of employees in each category is small, the point estimates indicate that lifestyle improvements by employees who were not sick drive large productivity gains. These gains are correlated with lifestyle improvements in stress, exercise, and HDL. It is apparent from these results that many employees without health problems made positive lifestyle changes due to the program, and these changes drove significant productivity growth for *LaundryCo*. Overall these results suggest capability improvements from improved overall and physical well-being.

We also examined whether improvements on specific health dimensions such as diabetes and kidney function were identifiable as driving the productivity gains in our main results. *Sick* and *Better* were defined by whether biometric test results were outside of or entered specified normal ranges, respectively. Although our sample size weakens inference on these small subgroups, the results (Appendix Figure A9) are supportive of the key mechanisms in Figure 5. The results show that improvements in disease areas associated with lifestyle choices, such as diabetes, electrolytes, or cholesterol, led to large productivity growth for sick individuals, although we caution that these data represent snapshots in blood tests that also may reflect random daily variation in blood chemistry.

### 4.3 Addressing Empirical Limitations

**4.3.1 Sample Size:** Despite its strengths, our setting has two weaknesses that should be addressed in future work. First, the number of workers for which we have two years of health data is only 111, which makes precise estimation of productivity changes for different subsamples difficult. Our small sample partly reflects *LaundryCo's* size, but also is hindered by high turnover that limits the length of time over which we can observe many workers. This problem is magnified by the small cell sizes for subgroup analysis. In addition to the 42 employees at Plant 1 serving as a control group, 55 of the 69 employees at the treatment plants chose to participate in the program. Of these 55, 36 qualified as sick and 15 improved their health. Only nine sick and six non-sick workers improved their health. Furthermore, the problem is particularly acute when attempting to identify the specific health improvements (e.g., diabetes) that drive productivity gains.

Consequently, the imprecision of our coefficients should not be interpreted as strong evidence of a null effect—we simply do not have enough statistical power to precisely identify small effect sizes.

Future work should seek organizations where larger samples will provide improved power and allow possible identification of smaller effects that support (or refute) our findings. Larger organizations would also allow for more detailed subsample analysis to pinpoint which specific health improvements are most crucial to improving capability and productivity. That our study is the first to link objective productivity and health data tells how difficult such a dataset would be to acquire. To help guide future research, we implement simulated power tests for our own sample size and then project how larger samples might improve statistical power. Power tests indicate the likelihood of a data sample correctly rejecting a null effect for a given parameter magnitude at a specific significance level ($\alpha$). To first test the power of our own sample, we randomly generate 1000 datasets, each with 52,293 observations distributed across pre- and post- blood draw periods for 111 workers: 42 are in the control group, 14 are non-compliers, 27 are sick non-improving compliers, 9 are sick improving compliers, and 6 are non-sick improving compliers. The productivity values for these observations are calculated using the parameter estimates in model (4) of Table 3 plus an error term based on two components: worker-level and individual error terms. The two error terms are normally-distributed random variables with mean zero and their respective standard deviations in our real sample, and weighted based on the calculated intra-cluster (worker) correlation in our data (0.41).[3]

The statistical power calculations using our sample size of 111 workers are provided in Table 4 for p-values of .05 and .10. As expected, statistical power is low, which helps explain the imprecision of our model estimates. Clearly, both our analysis and future studies would benefit from significantly increased sample size, although the non-existence of previous studies speaks to the difficulty of acquiring such samples.

<<< INSERT TABLE 4 HERE >>>

To provide guidance for future researchers, we repeat our simulation with different sample sizes, mapping the relationship between sample size and statistical power given our real sample's parameter estimates, cell-size ratios, and error structure. These results, presented in Figure 6, show that firm samples with greater than 500 long-term employees are likely necessary to reliably estimate precise effect sizes similar to ours. Smaller effects, such as may be the case for healthy, non-improving participants, would require even larger samples.

<<< INSERT FIGURE 6 HERE >>>

**4.3.2 Addressing endogenous participation:** The second weakness is the endogeneity of wellness program participation within the four participation plants as well as the choice to actively try to improve health. It could be that employees who agreed to participate were more likely to view the program in a positive light, and more likely to commit to lifestyle changes. Indeed, recent work has shown that individual worker differences such as time discounting or self-control can predict both health and other behavioral dimensions (Gubler and Pierce 2014; Israel et al. 2014). Although these results suggest the potential for instruments to

[3] These simulations were programmed in Stata with guidance from McConnell and Vera-Hernández (2015).

address endogenous improvement, such instruments would put further pressure on the statistical power issues discussed in the section above and are infeasible to implement here.

To address endogenous participation, we use local average treatment effects (LATE) models that estimate the average causal effect on compliers by instrumenting the random assignment of intended treatment status on endogenous compliance (Imbens and Angrist 1994). To implement this, we use a dummy variable *Post*TreatmentPlant* as an instrument for *Post*Complier* in two-stage least squares models with individual fixed effects that parallel models (1) and (2) in Table 3. The results, presented in Table A7 in the Appendix, show unbiased parameter estimates that are consistent with our model, but expectedly less precise.

**4.3.3 Attendance award program at control plant:** The control plant started an attendance award program in March 2011, nearly a full year into our study timeframe. Because it affected the control group only, its introduction might bias the results of our study if the award program reduced productivity in the control plant. Specifically, it might make the counterfactual time trend against which treatment group changes were compared appear worse than they otherwise would have been. Alternatively, if the award program had a positive effect on control group productivity, it would bias our study against finding an effect.

There are several reasons why the award program is unlikely to affect our results. First, research on the award program found no overall effect on productivity, making it is unlikely that the award program at the control plant is causing the results for the wellness program. The negative effect of the award program was only for a small sub-sample of employees who had strong attendance behavior before the award was implemented (Gubler et al. 2016). Second, if our results came from a deterioration in control group productivity, we would not find heterogeneous treatment effects across treatment group categories, since all control group members were assumed to be "*Not Sick, Not Better*" in our empirical design. That our empirical results did find the heterogeneous treatment effects predicted by our study suggests that the results are driven by treatment group changes, not control group changes. Indeed, raw plots of the data from Appendix Figure A5 do not show a drop off in control group productivity, and do show positive jumps in productivity for some (but not all) of the treatment subgroups.

## 5. Discussion and Conclusion

In this paper, we presented mechanisms through which firms might increase productivity by introducing formal programs that help employees track and improve health and wellness. We explained why the program's effects on employee productivity might depend on both an employee's pre-existing level of sickness and their post-program improvements in health. Notably, we explained that firms should not simply focus on enabling sick employees to identify and mitigate health concerns. Instead we argue that *all* types of employees might improve their productivity after the introduction of a wellness program.

While our study did not examine the long-term persistence of these effects, we did examine a full year in our lags and leads models. Although these models show that the large productivity gains from health improvers are unlikely to be generated through a temporary Hawthorne effect, the smaller and less precise gains from non-improving participants appear to dissipate after about 9 months, as seen in Figure 3. Thus, the motivational benefits from demonstrated commitment to workers may be short-lived, while the capability-based benefits from health improvement may be more persistent. Still, we caution that some of the persistent productivity gains we observe result from continued support by the firm for employee wellness. We doubt that long-term gains would be achieved through a single or short-term intervention.

Our empirical setting has several unique characteristics that make it a natural laboratory for studying employee health and productivity. First, our quasi-experimental setting and methods provide causal evidence that builds on previous work that was almost all correlational. Second, our health improvement data include both detailed objective medical tests as well as self-reported data. In addition, we used physician health evaluations to determine true health improvements beyond the objective normal ranges for blood tests. Third, our paper is the first to use objective productivity data, an important advance over biased self-reported productivity measures. But as we detail above, the small worker sample size makes our estimates imprecise, and they should be viewed as preliminary evidence. Future researchers would ideally study organizations with thousands of participants in order to achieve the statistical power necessary to precisely identify effect sizes similar to ours. We also note that such large samples would better allow researchers to dissect which specific health improvements are crucial for productivity gains. Our results in Figure 5 suggest that diet and exercise play key roles, but our small sample size limits our confidence in these conclusions.

### 5.1 Managerial Implications

Our empirical results demonstrate that the introduction of a corporate wellness program can have a large impact on employee productivity, and therefore firm profitability. Yet our study also suggests that the ability of an employer to enjoy a productivity-based return on investment (ROI) from the program crucially depends on two factors: the participation rate and employee turnover.

To estimate the ROI of the program, we first worked with senior executives at *LaundryCo* to estimate the value of a "free" hour of labor to the company; the estimate was that the labor itself was worth $15, while the more efficient use of its fixed capital stock (i.e., spreading fixed costs over more productivity) was worth $9. Therefore, an hour of "free" labor is worth $24 to the company. As seen in model (1) of Table 3, the average complier saw an increase in productivity of about 4%, meaning they worked nearly an additional hour of work per month due to the program. Assuming 220 days of work a year, and applying the coefficient estimate on compliers, this increase in productivity is worth $1,690 per complier. Given that there were 55 compliers, the benefit of the program for compliers was nearly $91,000. Non-compliers saw a productivity decrease of nearly 6%; the same logic means that the non-complier productivity drop cost the company $2,528 per non-complier. Since there were 14 non-compliers, the productivity decrease cost the company about $35,000. The total productivity benefit is the difference between these two numbers, or $57,558. We do note that both the estimates for compliers and non-compliers use imprecise coefficients from model (1), with p-values around

0.2. Furthermore, to be conservative, we have accounted for the full productivity decrease of non-compliers, even though the results suggest their productivity was already declining before the program was announced.

We next calculated the cost of the program. *LaundryCo* paid $120 per employee to the vendor for the program, and estimates that it spent another $120 per employee in terms of lost work time (to take the tests, have the follow-up visits, etc.). The program therefore cost $240 per participating worker. *LaundryCo* paid for 134 workers to take part in the program, 81 of whom left before they could participate a second time. This means that the total cost of the program was $32,640. The total ROI of the program was therefore 76.3%.

It is striking that the ROI on the program is large, despite the fact that turnover is so high and that conservative assumptions were used in its calculations. In the lower bound of our ROI estimates, *LaundryCo* is paying the full cost of the program for the large majority of participants, and is enjoying no productivity benefit from these participants; still, the ROI in this scenario is over 75%. Notably, the ROI is still relatively high even though slightly over 20% of workers choose not to participate, and suffer from a costly decrease in productivity, at least part of which was likely unrelated to the wellness program. Our analysis allows us to estimate the returns to a hypothetical "perfect" program in a company without turnover and where all employees chose to participate. In this case, the ROI would be nearly 590%.

Our results suggest even larger productivity gains for those whose health improved due to the program. Notably, there is no difference in the productivity growth for health-improving employees who were and were not identified as "sick" by the program. Again, this suggests that the impact of the program is more widespread than one might initially think, improving the capability of workers across the health spectrum. Indeed, survey results indicate that the wellness program led to lifestyle changes for employees regardless of sickness levels, likely generating capability improvements in both sick and healthy workers.

The results also suggest some caution, as one group of employees – those whom the program identified as sick but whose health did not subsequently improve – did not exhibit the productivity gains seen by the other three groups. Consequently, we see no evidence of gratitude-based reciprocity. Additionally, either they did not enjoy improved job satisfaction, or else such gains were cancelled out by the negative informational shock about and/or subsequent treatment for the indicated disease. Some employees may receive truly devastating news through a corporate wellness program, such as the existence of a terminal condition. This is unlikely to be the case in our empirical setting, where the most serious sicknesses uncovered by the program involved long-term manageable health conditions such as severe diabetes, obesity, and pain, not terminal diseases. Thus, we cannot conclude what combination of mechanisms might have generated this non-result.

### 5.2 Boundary Conditions

Although our empirical setting demonstrates how wellness programs can improve productivity, two important program design elements define boundary conditions for such improvements. First, employee participation cannot be compulsory or heavily coerced through social pressure or financial penalties that might induce psychological reactance. Psychological reactance theory (Brehm 1966) argues that individuals strongly react to external influence that they perceive to restrict their autonomy. Programs such as wellness initiatives that threaten worker autonomy might motivate employees to assert their autonomy either through resisting the program or even through reduced productivity. Indeed, public health scholars argue that strong incentives and requirements in wellness programs can produce negative effects through psychological reactance (Dowd 2002), since employees view health and lifestyle choices as outside their work domain. The program in our empirical setting was both voluntary and only weakly incentivized.

Second, employees must trust that the firm will respect the privacy of employee health data and not use it for employment-related purposes. HIPAA regulations in the United States forbid firms from accessing employee health and wellness data collected through group health plans. However, data from employer-run wellness programs may be legally accessible. While firms cannot legally use these data for employment decisions, and must formally separate program administration from other human resource functions, employees may not trust the firm to observe this prohibition. Firms must not only observe these regulations, but also communicate and demonstrate this compliance for credibility with employees. Employees who mistrust the firm's use of private health data might view the wellness program as violating a broader psychological contract that governs their overall relationship and influences their individual day-to-day actions (Rousseau 1990). This perceived abrogation by the employer of a part of this implicit contract could reduce overall job motivation, satisfaction, and retention among those who strongly value health privacy.

The imprecise negative effect estimate on non-compliers could reflect perceptions from these select employees that the program threatened either their autonomy or privacy, thereby reducing their future motivation despite not participating. This highlights the important difference between the reality of the program and its perception by workers. Even the best-designed wellness programs may threaten a small group of employees, and the firm must actively convince these workers of the program's merits and safety.

## References

Adams G, Flynn F, Norton M (2012) The gifts we keep on giving: Documenting and destigmatizing the regifting taboo. *Psychological Science,* 23(10): 1145-1150.

Agle BR, Mitchell RK, Sonnenfeld JA (1999) Who matters to CEOs? An investigation of stakeholder attributes and salience, corporate performance, and CEO values. *Academy of Management Journal,* 42(5): 507–525.

Aguilera RV, Rupp DE, Williams CA, Ganapathi J (2007) Putting the S back in corporate social responsibility: A multilevel theory of social change in organizations. *Academy of Management Review,* 32(3): 836–863.

Armeli S, Eisenberger R, Fasolo P, Lynch P (1998) Perceived organizational support and police performance: The moderating influence of socioemotional needs. *Journal of Applied Psychology,* 83(2): 288.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Autor D (2003) Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of Labor Economics 21*(1): 1-42.

Baicker K, Cutler D, Song Z (2010) Workplace wellness programs can generate savings. *Health Affairs,* 29(2): 304-311.

Barnett ML, Salomon RM (2006) Beyond dichotomy: The curvilinear relationship between social responsibility and financial performance. *Strategic Management Journal,* 27(11): 1101–1122.

Bartlett MY, DeSteno D (2006) Gratitude and prosocial behavior helping when it costs you. *Psychological Science,* 17(4): 319-325.

Beam CA, Layde PM, Sullivan DC (1996) Variability in the interpretation of screening mammograms by US radiologists. *Archives of Internal Medicine,* 156: 209-213.

Becker GS (2007) Health as human capital: synthesis and extensions. *Oxford Economic Papers,* 59(3): 379–410.

Bernstein ES (2012) The transparency paradox a role for privacy in organizational learning and operational control. *Administrative Science Quarterly 57*(2): 181-216.

Berry L, Mirabito AM, Baun WB (2010) What's the hard return on employee wellness programs? *Harvard Business Review,* December: 2012-68.

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

Beshears J, Milkman K, Schwartzstein J (2016) Beyond beta-delta: The emerging economics of personal plans. *The American Economic Review P&P*, *106*(5), 430-434.

Boles M, Pelletier B, Lynch W (2004) The relationship between health risks and work productivity. *Journal of Occupational and Environmental Medicine*, 46(7): 737-745.

Brehm JW (1966) *A Theory of Psychological Reactance* (Academic Press, New York).

Buell RW, Kim T, Tsay CJ (2017) Creating reciprocal value through operational transparency. *Management Science* 63(6): 1673-1695.

Burbano V (2017a) Social responsibility messages and worker wage requirements: Field experimental evidence from online labor marketplaces. *Organization Science*. Forthcoming.

Burbano V (2017b) Getting gig workers to do more by doing good: Field experimental evidence from online platform labor marketplaces. *Unpublished working paper.*

Burton WN, Chen CY, Conti DJ, Schultz AB, Pransky G, Edington DE (2005) The association of health risks with on-the-job productivity. *Journal of Occupational and Environmental Medicine,* 14(8): 767–777.

Calzolari G, Nardotto M (2017) Effective reminders. *Management Science.* Forthcoming

Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2): 317-372.

Carmeli A, Gilat G, Waldman DA (2007) The role of perceived organizational performance in organizational identification, adjustment and job performance. *Journal of Management Studies,* 44(6): 972–992.

Carnahan S, Kryscynski D, Olson D (2017) How corporate social responsibility reduces employee turnover: Evidence from attorneys before and after 9/11. *Academy of Management Journal.* Forthcoming.

Center for Disease Control (2014) *Diabetes Public Health Resource.* Available at: http://www.cdc.gov/diabetes/statistics/incidence/fig1.htm

Chan TY, Li J, Pierce L (2014a) Compensation and peer effects in competing sales teams. *Management Science 60*(8): 1965-1984.

Chan TY, Li J, Pierce L (2014b) Learning from peers: Knowledge transfer and sales force productivity growth. *Marketing Science* 33(4): 463-484.

Chapman LS (2012) Meta-evaluation of worksite health promotion economic return studies: 2012 update. *American Journal of Health Promotion,* 26(4): TAHP-1.

Chatterji AK, Levine DI, Toffel MW (2009) How well do social ratings actually measure corporate social responsibility? *Journal of Economics & Management Strategy*, 18: 125-169.

Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT (2007) Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Archives Ophthalmology,* 125(7): 875-880.

Christian MS, Eisenkraft N, Kapadia C (2015) Dynamic associations among somatic complaints, human energy, and discretionary behaviors experiences with pain fluctuations at work. *Administrative Science Quarterly,* 60(1): 66-102.

Conti G, Heckman J, Urzua S (2010) The education-health gradient. *American Economic Review,* 100(2): 234–238.

Currie J, Madrian BC (1999) Health, health insurance and the labor market: Chapter 50. In Orley C. Ashenfelter and David Card, ed. *Handbook of Labor Economics.* Elsevier, 3309–3416.

Dabos GE, Rousseau DM (2004) Mutuality and reciprocity in the psychological contracts of employees and employers. *Journal of Applied Psychology*, 89(1): 52.

Dai H, Milkman K, Beshears J, Choi J, Laibson D, Madrian B (2012). Planning prompts as a means of increasing rates of immunization and preventive screening. *Public Policy & Aging Report* 22(4): 16-19.

Dai H, Milkman KL, Hofmann DA, Staats BR (2015) The impact of time at work and time off from work on rule compliance: The case of hand hygiene in health care. *Journal of Applied Psychology 100*(3): 846.

Danna K, Griffin RW (1999) Health and well-being in the workplace: A review and synthesis of the literature. *Journal of Management,* 25(3): 357-384.

Dowd E (2002) Psychological reactance in health education and promotion. *Health Education Journal* 61(2): 113-124.

Eddy D, Clanton C (1992) The art of diagnosis: solving the clinicopathological exercise. *New England Journal of Medicine,* 306(21): 1263-1268.

Einhorn HJ (1972) Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, 7:86-106.

Eisenberger R, Armeli S, Rexwinkel B, Lynch PD, Rhoades L (2001) Reciprocation of perceived organizational support. *Journal of Applied Psychology,* 86(1): 42.

Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR (1994) Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine,* 331(22): 1493-1499.

Flammer C (2015) Does corporate social responsibility lead to superior financial performance? A regression discontinuity approach. *Management Science,* 61(11): 2549-2568.

Flammer C, Luo J (2017) Corporate social responsibility as an employee governance tool: Evidence from a quasi-experiment. *Strategic Management Journal* 38(2): 163-183.

Gallup-Healthways (2012) *Well-being Index*. Available at: http://www.healthways.com/solution/default.aspx?id=1125 [accessed December 9, 2012]

Gertler PJ, Martinez S, Premand P, Rawlings LB, Varmeersch CMJ (2011) *Impact evaluation in practice* (World Bank Publications).

Glavas A, Piderit SK (2009) How does doing good matter? Effects of corporate citizenship on employees. *Journal of Corporate Citizenship,* 36: 51–70.

Goetzel RZ, Hawkins K, Ozminkowski RJ, Wang S (2003) The health and productivity cost burden of the "Top 10" physical and mental health conditions affecting six large U.S. employers in 1999. *Journal of Occupational and Environmental Medicine,* 45(1): 5–14.

Goldstein SM (2003) Employee development: an examination of service strategy in a high-contact service environment. *Production and Operations Management 12*(2): 186-203.

Grant AM, Gino F (2010) A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior. *Journal of Personality and Social Psychology,* 98(6): 946.

Grant AM, Christianson M, Price R (2007) Happiness, health, or relationships? Managerial practices and employee well-being tradeoffs. *Academy of Management Perspectives* 21(3): 51-63.

Gubler T, Larkin I, Pierce L (2016) Motivational spillovers from awards: Crowding out in a multitasking environment. *Organization Science 27*(2): 286-303.

Gubler T, Pierce L (2014) Healthy, wealthy, and wise: Retirement planning predicts employee health improvements. *Psychological Science*, 25(9): 1822-1830.

Hekman DR, Bigley GA, Steensma HK, Hereford JF (2009) Combined effects of organizational and professional identification on the reciprocity dynamic for professional employees. *Academy of Management Journal,* 52(3): 506-526.

Huckman RS, Pisano GP (2006). The firm specificity of individual performance. Evidence from cardiac surgery. *Management Science* 52(4): 473-488.

Huckman RS, Staats BR (2011). Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance. *Manufacturing & Service Operations Management 13*(3): 310-328.

Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. *Econometrica*, 62(2): 467-475.

Ingraham C (2016) Nearly half of America's overweight people don't realize they're overweight. *Washington Post*. December 1, 2016.

Israel S, Caspi A, Belsky D, Harrington H, Hogan S, Houts R, Ramrakha S, Sanders S, Poulton R, Moffitt T (2014) Credit scores, cardiovascular disease risk, and human capital. *Proceedings of the National Academy of Sciences,* 111(48): 17087-17092.

Jones DA (2010) Does serving the community also serve the company? Using organizational identification and social exchange theories to understand employee responses to a volunteerism programme. *Journal of Occupational and Organizational Psychology,* 83(4): 857–878.

Kaiser/HRET (2012) *Survey of Employer-Sponsored Health Benefits, 1999-2009.* Available at: http://www.hret.org/reform/projects/employer-health-benefits-annual-survey.shtml

KC DS (2013) Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management 16*(2): 168-183.

Kitzmueller M, Shimshack J (2012) Economic perspectives on corporate social responsibility. *Journal of Economic Literature,* 50(1): 51–84.

Kroboth F, Hanusa B, Parker S, Coulehan J, Kapoor W, Brown F, Karpf M, Levey S (1992) The inter-rater reliability and internal consistency of a clinical evaluation exercise. *Journal of General Internal Medicine* 7(2): 174-179.

Kuntz L, Mennicken R, Scholtes S (2014) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science 61*(4): 754-771.

Larkin, I (2011) Paying 30,000 for a gold star: An empirical investigation into the value of peer recognition to software salespeople, working paper, Harvard Business School.

Lee JJ, Gino F, Staats BR (2014) Rainmakers: Why bad weather means good productivity. *Journal of Applied Psychology*, *99*(3): 504.

Margolis JD, Walsh JP (2003) Misery loves companies: rethinking social initiatives by business. *Administrative Science Quarterly,* 48: 268-305.

McConnell B, Vera-Hernandez M (2015) Going beyond simple sample size calculations: A practitioner's guide. *Institute for Fiscal Studies Working Paper.* (No. W15/17).

Medical Billing and Coding (2012) Available at: http://www.medicalbillingandcoding.org/blog/12-companies-with-seriously-impressive-corporate-wellness-programs/

Mensink RP, Zock PL, Kester ADM, Katan MB (2003) Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: a meta-analysis of 60 controlled trials. *American Journal of Clinical Nutrition,* 77(5): 1146-1155.

Milkman K, Beshears J, Choi J, Laibson D., Madrian B (2011) Using implementation intentions prompts to enhance influenza vaccination rates. *Proceedings of the National Academy of Sciences*, *108*(26): 10415-10420.

NCHS (2012) *NCHS Data Brief No. 82.* Available at: http://www.cdc.gov/nchs/data/databriefs/db82.pdf

Neumann W, Dul J (2010) Human factors: spanning the gap between OM and HRM. *International Journal of Operations & Production Management 30*(9): 923-950.

Ødegaard F, Roos P (2014) Measuring the Contribution of Workers' Health and Psychosocial Work-Environment on Production Efficiency. *Production and Operations Management 23*(12): 2191-2208.

Orlitzky M, Schmidt FL, Rynes SL (2003) Corporate social and financial performance: A meta-analysis. *Organization Studies,* 24: 403-441.

Parks KM, Steelman LA (2008) Organizational wellness programs: A meta-analysis. *Journal of Occupational Health Psychology,* 13(1): 58.

Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP (2003) Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology,* 88(5): 879.

Rogers T, Milkman KL, John LK, Norton MI (2015) Beyond good intentions: Prompting people to make plans improves follow-through on important tasks. *Behavioral Science & Policy*, *1*(2): 33-41.

Rousseau DM (1990) New hire perceptions of their own and their employer's obligations: A study of psychological contracts. *Journal of Organizational Behavior,* 11: 389-400.

Song H, Tucker A, Murrell K, Vinson D (2017) Public relative performance feedback in complex service systems: Improving productivity through the adoption of best practices. *Management Science.* Forthcoming

Staats B, Dai H, Hofmann D, Milkman K (2016) Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Management Science*. Forthcoming

Staats B, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science 58*(6): 1141-1159.

Stewart WF, Ricci JA, Chee E, Morganstein D (2003) Lost productive work time costs from health conditions in the United States: Results from the American productivity audit. *Journal of Occupational and Environmental Medicine,* 45(12): 1234-1246.

Tan TF, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* 60(6): 1574-1593.

Tan TF, Netessine S (2015) When You Work with a Super Man, Will You Also Fly? An Empirical Study of the Impact of Coworkers on Performance. Working Paper.

Thayer RE, Newman JR, McClain TM (1994) Self-regulation of mood: strategies for changing a bad mood, raising energy, and reducing tension. *Journal of Personality and Social Psychology,* 67(5): 910.

Ton Z. (2014). *The good jobs strategy: How the smartest companies invest in employees to lower costs and boost profits*. Houghton Mifflin Harcourt.

Tsang JA (2006) Gratitude and prosocial behaviour: An experimental test of gratitude. *Cognition & Emotion*, 20(1): 138-148.

Turban DB, Greening DW (1997) Corporate social performance and organizational attractiveness to prospective employees. *Academy of Management Journal,* 40(3): 658–672.

Vallgårda S (2012) Nudge: A new and better way to improve health? *Health Policy,* 104(2):200-203.

Wang H, Qian C (2011) Corporate philanthropy and corporate financial performance: The roles of stakeholder response and political access. *Academy of Management Journal.* 54(6):1159–1181.

Ward, B, Schiller J and Goodman R (2014) Multiple chronic conditions among US adults: A 2012 update. *Preventing Chronic Disease.* 11(April): 1-4.

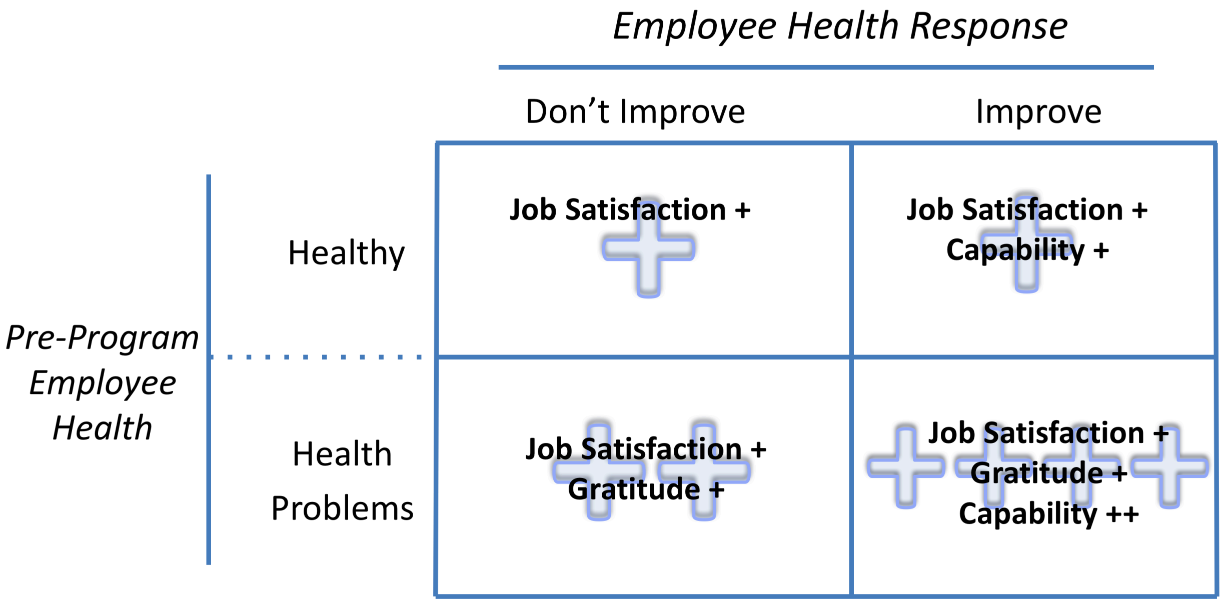
Weber L (2014) Wellness programs get a health check, *The Wall Street Journal*, Available at: http://www.wsj.com/articles/wellness-programs-get-a-health-check-1412725776

Wood PD, Stefanick ML, Dreon DM, Frey-Hewitt B, Garay SC, Williams PT, Superko HR, Formann SP, Albers JJ, Vranizan KM, Ellsworth NM, Terry RB, Haskell WL (1988) Changes in plasma lipids and lipoproteins in overweight men during weight loss through dieting as compared with exercise. *The New England Journal of Medicine*, 319(18): 1173-1179.

Yee RW, Yeung AC, Cheng TE (2008) The impact of employee satisfaction on quality and profitability in high-contact service industries. *Journal of Operations Management* 26(5): 651-668.

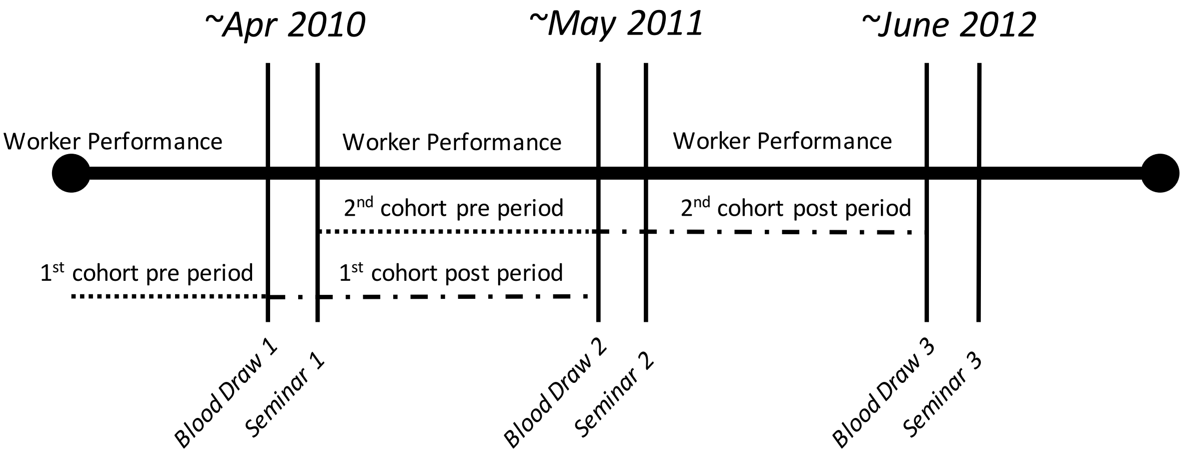
Zoller HM (2004) Manufacturing health: Employee perspectives on problematic outcomes in a workplace health promotion initiative. *Western Journal of Communication* 68(3): 278-301.

Figure 1: Four Participant Types and Possible Mechanisms for Productivity Improvement



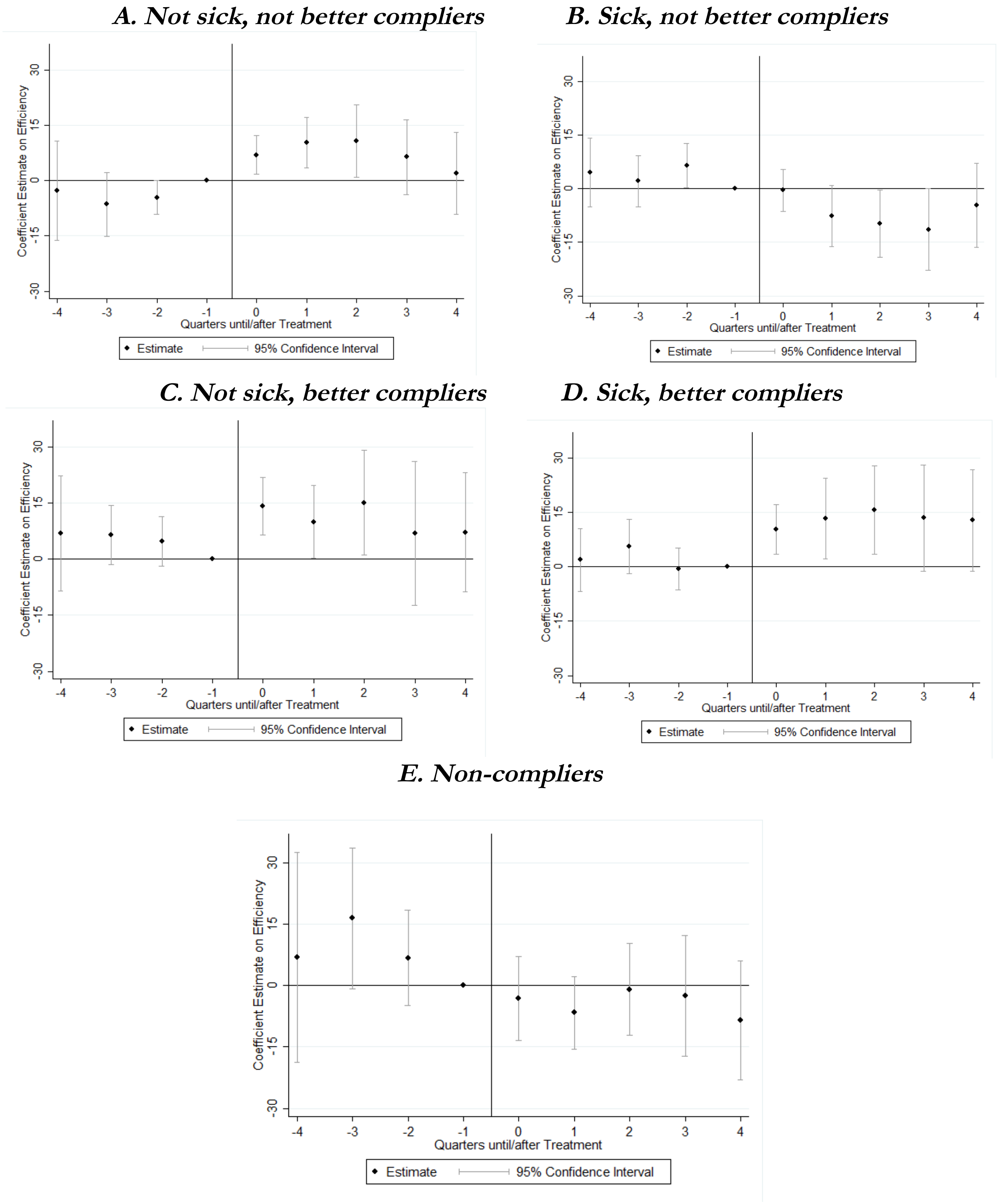
|  |  | *Employee Health Response*<br>Don't Improve | *Employee Health Response*<br>Improve |
|---|---|---|---|
| *Pre-Program Employee Health* | Healthy | **Job Satisfaction +** | **Job Satisfaction +<br>Capability +** |
|  | Health Problems | **Job Satisfaction +<br>Gratitude +** | **Job Satisfaction +<br>Gratitude +<br>Capability ++** |

Figure 2: Timing of Wellness Program



Note: Blood draw dates refer to when blood was drawn and surveys filled out at each plant. Seminar dates refer to when health information was returned in a personalized packet to each participant. This typically occurred 3-4 weeks post blood draw. We only use data from the first time participating for each worker. The 2010 pre-period begins in 2009.

Figure 3: Lead and Lags Results


A. Not sick, not better compliers

B. Sick, not better compliers

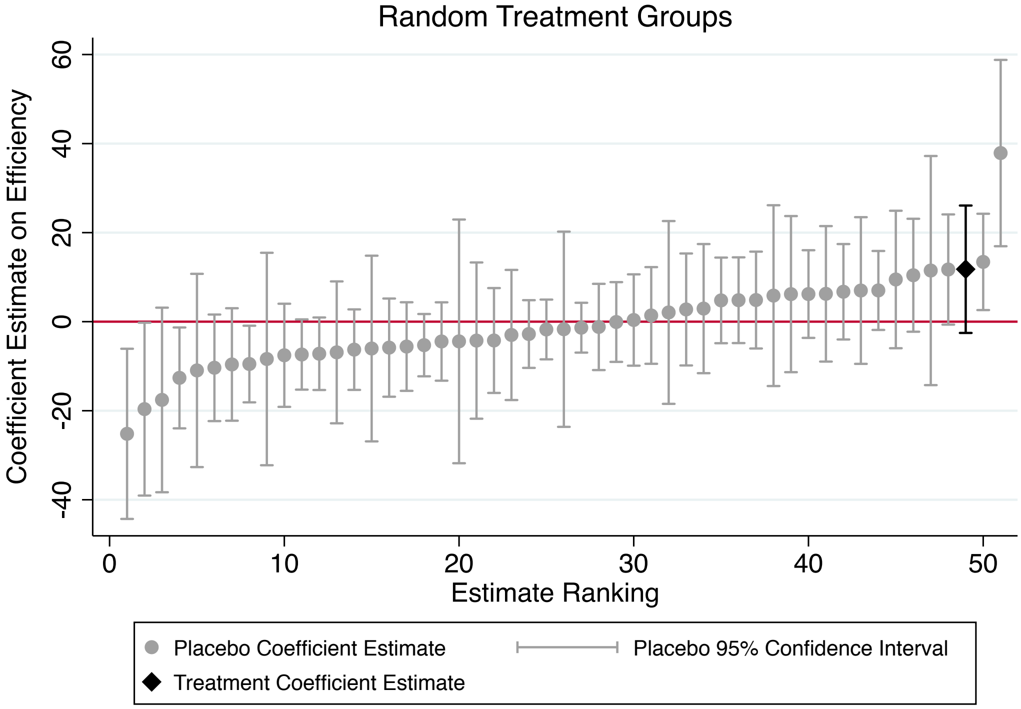
C. Not sick, better compliers

D. Sick, better compliers

E. Non-compliers

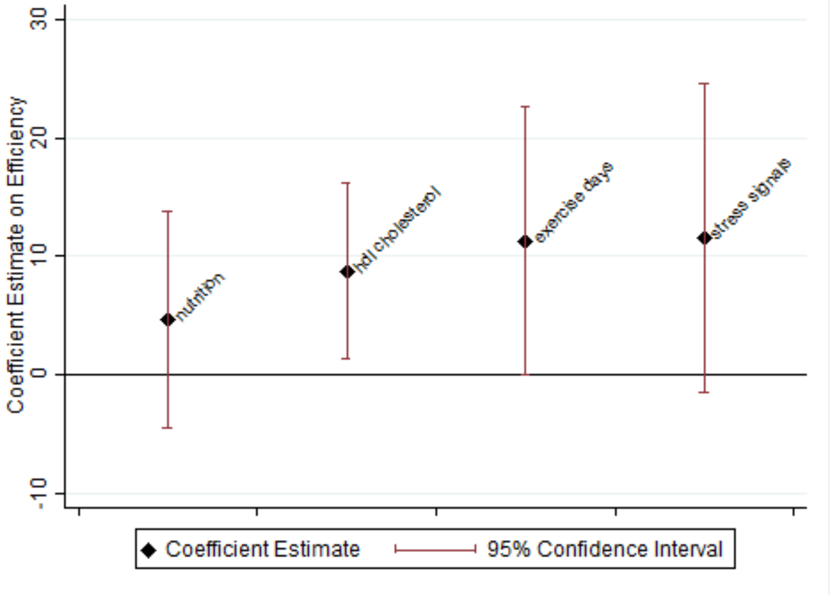
Note: Dots represent estimate of difference between treatment and control groups compared to difference in Bucket -1. Bucket 0 contains the program announcement, blood draw and seminar. Bucket -1, the omitted interaction, contains the three months just before program announcement.

## Figure 4: Placebo Test on Sick, Better Employee Group



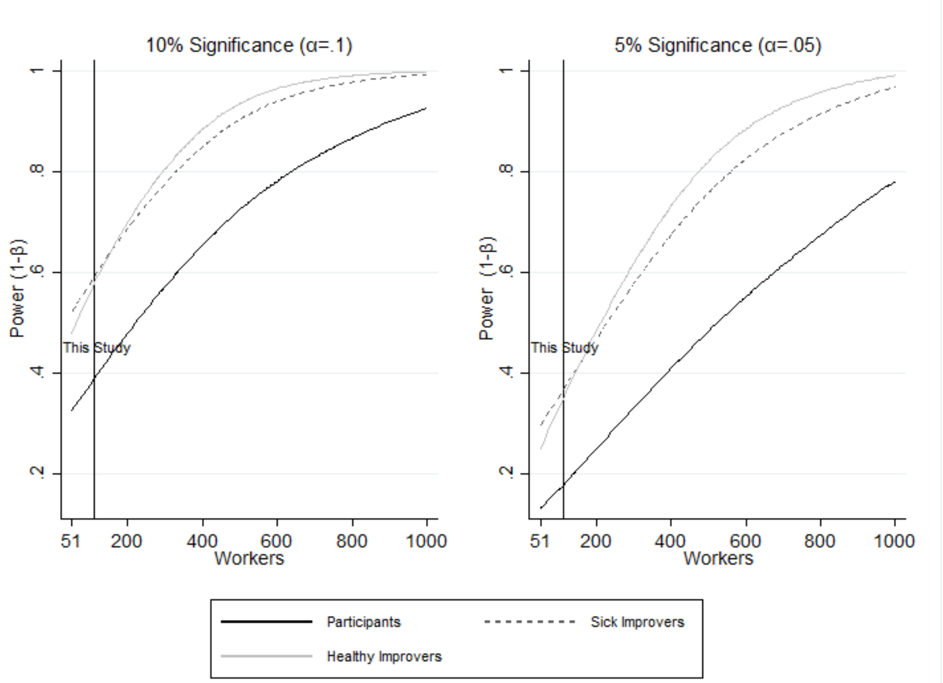
Note: 50 placebo simulations for the main model using randomized treatment groups and treatment dates.

## Figure 5: Regression Coefficients Using Survey Measures



Note: Diamonds represent coefficient estimates for each specific category following the main specification. Models are estimated using OLS with clustered standard errors at the individual level. "Sick" is defined as having a nutrition score <50, hdl cholesterol <=39, not exercising, and reporting a stress signal. "Better" is defined as positive improvements on a given category. Coefficient estimates are plotted for sick individuals that improve ("Better" equal to 1). Estimates based off 14 improvers in stress, 12 in nutrition, 21 in exercise, and 37 in HDL.

## Figure 6: Projected Statistical Power for Larger Worker Samples



Note: These two figures represent estimated statistical power for hypothetical different sample sizes for our main model and its parameter estimates in Table 3. Sample sizes were increased in intervals of 10, from 51 workers to 1001 workers while maintaining the same proportions of compliers, non-compliers, sick, and improving people. For each sample size, the power estimate is based on 1000 simulated datasets. The vertical line at 111 workers represents our true sample.

## Table 1: Descriptive Statistics at Worker/Day Level

| Variable | Count | Mean | Sd | Min | Max |
|---|---|---|---|---|---|
| Efficiency | 52293 | 125.83 | 31.60 | 24.67 | 264.10 |
| Post period | 52293 | 0.47 | 0.50 | 0 | 1 |
| Month(continuous) | 52293 | 17.64 | 10.05 | 0 | 42 |
| Year | 52293 | 2009 | 0.87 | 2009 | 2012 |
| Participant | 52293 | 0.52 | 0.50 | 0 | 1 |
| Sick (doctor measure) | 52293 | 0.33 | 0.47 | 0 | 1 |
| Better (doctor measure) | 52293 | 0.14 | 0.35 | 0 | 1 |
| Better (doctor group measure) | 52293 | 0.17 | 0.37 | 0 | 1 |
| Better exercise days | 52293 | 0.21 | 0.41 | 0 | 1 |
| Better hdl | 52293 | 0.35 | 0.48 | 0 | 1 |
| Better nutrition score | 52293 | 0.12 | 0.33 | 0 | 1 |
| Better stress signals | 52293 | 0.13 | 0.33 | 0 | 1 |
| N | 52293 | | | | |

## Table 2: Correlations for Primary Variables

| | | Post | | | | Better | Better | Better | Better | Better |
|---|---|---|---|---|---|---|---|---|---|---|
| | Efficiency | period | Participant | Sick | Better | (group) | exercise | hdl | nutrition | stress |
| Efficiency | 1.00 | | | | | | | | | |
| Post period | 0.04 | 1.00 | | | | | | | | |
| Participant | 0.03 | 0.03 | 1.00 | | | | | | | |
| Sick | -0.13 | 0.04 | 0.67 | 1.00 | | | | | | |
| Better | 0.00 | 0.02 | 0.39 | 0.20 | 1.00 | | | | | |
| Better (group) | -0.03 | 0.03 | 0.43 | 0.29 | 0.91 | 1.00 | | | | |
| Better exercise | 0.08 | -0.03 | 0.50 | 0.28 | 0.34 | 0.29 | 1.00 | | | |
| Better hdl | 0.03 | 0.03 | 0.71 | 0.52 | 0.19 | 0.23 | 0.40 | 1.00 | | |
| Better nutrition | -0.02 | 0.00 | 0.36 | 0.00 | 0.30 | 0.32 | 0.11 | 0.38 | 1.00 | |
| Better stress | -0.02 | 0.00 | 0.37 | 0.16 | 0.33 | 0.28 | 0.16 | 0.12 | 0.24 | 1.00 |

## Table 3: Regression Estimates (Total Effects)

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable | Efficiency | Efficiency | Efficiency | Efficiency |
| Model | OLS | OLS | OLS | OLS |
| Physician Scale | Independent Doctors | Independent Doctors | Independent Doctors | Independent Doctors |
| Non-compliers | -7.209 | -7.128 | -7.598 | -7.456 |
| | (5.331) [0.179] | (5.244) [0.177] | (5.403) [0.162] | (5.328) [0.165] |
| Compliers | 4.888 | | | |
| | (3.971) [0.229] | | | |
| Non-sick compliers | | 9.392* | | |
| | | (4.818) [0.054] | | |
| Sick compliers | | 0.979 | | |
| | | (4.004) [0.807] | | |
| Non-better compliers | | | 1.552 | |
| | | | (3.657) [0.672] | |
| Better compliers | | | 12.670** | |
| | | | (5.969) [0.036] | |
| Non-sick, non-better compliers | | | | 7.662* |
| | | | | (4.321) [0.079] |
| Sick, non-better compliers | | | | -3.068 |
| | | | | (4.213) [0.468] |
| Non-sick, better compliers | | | | 13.461* |
| | | | | (7.423) [0.073] |
| Sick, better compliers | | | | 11.778 |
| | | | | (7.220) [0.106] |
| Control for worker experience | Y | Y | Y | Y |
| Plant time trends | Y | Y | Y | Y |
| Fixed effects | Y | Y | Y | Y |
| $R^2$ | 0.455 | 0.458 | 0.463 | 0.465 |
| # of employees | 111 | 111 | 111 | 111 |
| Observations | 52293 | 52293 | 52293 | 52293 |
| Sick cutoff | 3 doctors | 3 doctors | 3 doctors | 3 doctors |

Note: Robust standard errors in parentheses, clustered by individual. P-values in brackets. The dependent variable is daily worker efficiency. All three physicians must specify an individual as "sick" for "sick" to take the value of 1. Results are robust to a cutoff of 2 physicians specifying sickness. "Better" uses the average of the physician's evaluations on improvement (Q4 in the Physician Evaluation Questionnaire), and takes the value of 1 if the average indicates improvement (>3 on the Q4 5-point scale). ** p<.05
* p<0.10

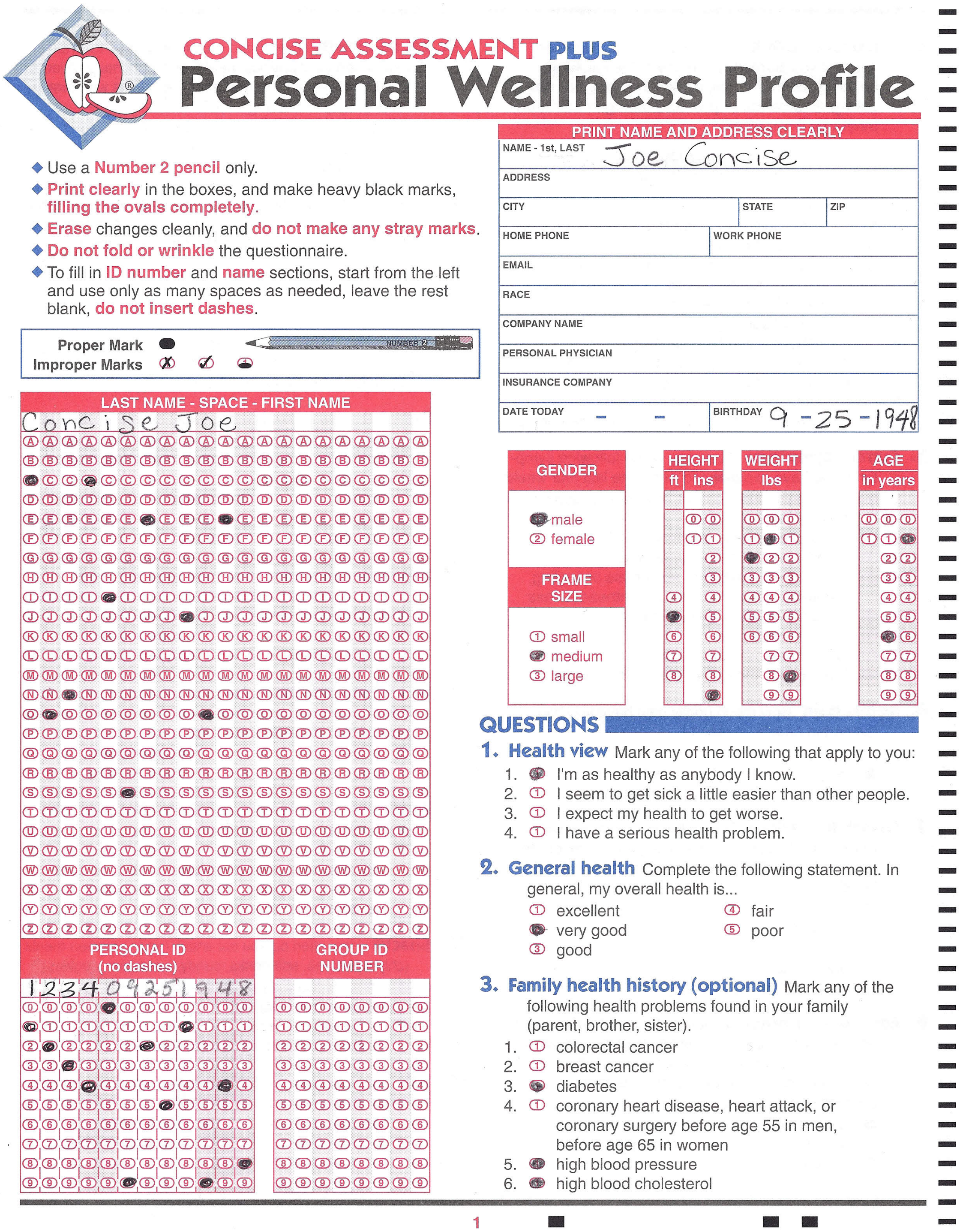
## Table 4: Simulated Power Tests for 111 Worker Sample

| | | (1) | (2) |
|---|---|---|---|
| | Dependent Variable | Efficiency | Efficiency |
| | Model | OLS | OLS |
| Cell Size<br>(Workers) | Health Assessment | Independent Doctors | Independent Doctors |
| | P-Value | 0.05 | 0.1 |
| 13 | Participant x Post | 0.144 | 0.346 |
| 14 | Non-Participant x Post | 0.424 | 0.543 |
| 27 | Participant x Post x Sick | 0.209 | 0.460 |
| 6 | Participant x Post x Better | 0.335 | 0.553 |
| 9 | Participant x Post x Sick x Better | 0.321 | 0.667 |
| | # of employees | 111 | 111 |
| | # simulation iterations | 1000 | 1000 |
| | Observations | 52293 | 52293 |

Note: Values represent probability that null hypothesis of zero effect is rejected, given the total effects estimated in column (1) of Table 3 and the p-value for each column. Power values are based on 1000 simulated datasets with the real sample's data structure. Larger cell sizes and larger effect sizes produce higher power values. All simulations use 42 control workers.

# Online Appendix for "Doing Well by Making Well"

## Figure A1. WellSource Survey (with example/fictitious answers)



CONCISE ASSESSMENT PLUS
# Personal Wellness Profile

- Use a **Number 2 pencil** only.
- **Print clearly** in the boxes, and make heavy black marks, **filling the ovals completely**.
- **Erase** changes cleanly, and **do not make any stray marks**.
- **Do not fold or wrinkle** the questionnaire.
- To fill in **ID number** and **name** sections, start from the left and use only as many spaces as needed, leave the rest blank, **do not insert dashes**.

Proper Mark ●
Improper Marks ☒ ✓ ⊖

PRINT NAME AND ADDRESS CLEARLY

NAME - 1st, LAST Joe Concise
ADDRESS
CITY
STATE
ZIP
HOME PHONE
WORK PHONE
EMAIL
RACE
COMPANY NAME
PERSONAL PHYSICIAN
INSURANCE COMPANY
DATE TODAY – –
BIRTHDAY 9 -25 -1948

LAST NAME - SPACE - FIRST NAME

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | o | n | c | i | s | e | | J | o | e | | | | | | | | | |
| A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A |
| B | B | B | B | B | B | B | B | B | B | B | B | B | B | B | B | B | B | B | B |
| ● | C | C | ● | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C | C |
| D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D |
| E | E | E | E | E | E | ● | E | E | E | ● | E | E | E | E | E | E | E | E | E |
| F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F |
| G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G |
| H | H | H | H | H | H | H | H | H | H | H | H | H | H | H | H | H | H | H | H |
| I | I | I | I | ● | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I |
| J | J | J | J | J | J | J | J | ● | J | J | J | J | J | J | J | J | J | J | J |
| K | K | K | K | K | K | K | K | K | K | K | K | K | K | K | K | K | K | K | K |
| L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L |
| M | M | M | M | M | M | M | M | M | M | M | M | M | M | M | M | M | M | M | M |
| N | N | ● | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N |
| O | ● | O | O | O | O | O | O | O | ● | O | O | O | O | O | O | O | O | O | O |
| P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P |
| Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q | Q |
| R | R | R | R | R | R | R | R | R | R | R | R | R | R | R | R | R | R | R | R |
| S | S | S | S | S | ● | S | S | S | S | S | S | S | S | S | S | S | S | S | S |
| T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T | T |
| U | U | U | U | U | U | U | U | U | U | U | U | U | U | U | U | U | U | U | U |
| V | V | V | V | V | V | V | V | V | V | V | V | V | V | V | V | V | V | V | V |
| W | W | W | W | W | W | W | W | W | W | W | W | W | W | W | W | W | W | W | W |
| X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z | Z |

PERSONAL ID (no dashes)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 0 | 9 | 2 | 5 | 1 | 9 | 4 | 8 |
| 0 | 0 | 0 | 0 | ● | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ● | 1 | 1 | 1 | 1 | 1 | 1 | 1 | ● | 1 | 1 | 1 |
| 2 | ● | 2 | 2 | 2 | 2 | ● | 2 | 2 | 2 | 2 | 2 |
| 3 | 3 | ● | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 4 | 4 | 4 | ● | 4 | 4 | 4 | 4 | 4 | 4 | ● | 4 |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | ● | 5 | 5 | 5 | 5 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | ● |
| 9 | 9 | 9 | 9 | 9 | ● | 9 | 9 | 9 | ● | 9 | 9 |

GROUP ID NUMBER

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |

GENDER

● male
② female

FRAME SIZE

① small
● medium
③ large

| HEIGHT ft | ins | | WEIGHT lbs | | | AGE in years | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 1 | 1 | ● | 1 | 1 | 1 | ● |
| | | 2 | ● | 2 | 2 | | 2 | 2 |
| | | 3 | 3 | 3 | 3 | | 3 | 3 |
| 4 | | 4 | 4 | 4 | 4 | | 4 | 4 |
| ● | | 5 | 5 | 5 | 5 | | 5 | 5 |
| 6 | | 6 | 6 | 6 | 6 | | ● | 6 |
| 7 | | 7 | | 7 | 7 | | 7 | 7 |
| 8 | | 8 | | 8 | ● | | 8 | 8 |
| | | ● | | 9 | 9 | | 9 | 9 |

## QUESTIONS

**1. Health view** Mark any of the following that apply to you:

1. ● I'm as healthy as anybody I know.
2. ① I seem to get sick a little easier than other people.
3. ① I expect my health to get worse.
4. ① I have a serious health problem.

**2. General health** Complete the following statement. In general, my overall health is...

① excellent
● very good
③ good
④ fair
⑤ poor

**3. Family health history (optional)** Mark any of the following health problems found in your family (parent, brother, sister).

1. ① colorectal cancer
2. ① breast cancer
3. ● diabetes
4. ① coronary heart disease, heart attack, or coronary surgery before age 55 in men, before age 65 in women
5. ● high blood pressure
6. ● high blood cholesterol

1

## 4. Personal health history

Has a doctor informed you that you currently have any of the following health problems? **If yes**, mark either ***yes, but not taking medication*** or ***yes, and taking medication***, otherwise leave blank.

***1 - yes, but not taking medication***
***2 - yes, and taking medication***

1. ① ② asthma
2. ● ② bowel polyps or inflammatory bowel disease
3. ① ② cancer, other than skin cancer
4. ① ② chronic bronchitis or emphysema (COPD)
5. ① ② coronary heart disease, congestive heart failure, angina, heart attack, or heart surgery
6. ① ② diabetes (high blood sugar)
7. ① ● high blood pressure (140/90 or higher)
8. ① ● high blood cholesterol (240 or higher)
9. ① ② sciatica or chronic back problem
10. ① ② stroke or restricted blood flow to head or legs

## 5. Current symptoms

Mark any of the following symptoms you have experienced within the last four weeks.

1. ① chest pain or discomfort, frequent palpitations or fluttering in the heart
2. ① unusual shortness of breath
3. ① unexplained dizziness or fainting
4. ① temporary sensation of numbness or tingling, paralysis, vision problem, or lightheadedness
5. ① frequent urination and unusual thirst
6. ① frequent back pain
7. ① have trouble sleeping lately
8. ① I've recently thought about ending my life

## 6. Bodily pain

How much bodily pain have you had during the past four weeks?

① none
● very mild
③ mild
④ moderate
⑤ severe
⑥ very severe

## 7. Health limitations

During the past four weeks, how much difficulty did you have doing your work or other regular daily activities as a result of your physical health?

● none
② a little bit
③ some
④ quite a bit
⑤ could not do daily work

## 8. Emotional problems

During the past four weeks, to what extent have you accomplished less than you would like in your work or other daily activities as a result of emotional problems, such as feeling depressed or anxious?

● none at all
② slightly
③ moderately
④ quite a bit
⑤ extremely

## 9. Social activity

During the past four weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

● none at all
② slightly
③ moderately
④ quite a bit
⑤ extremely

## 10. Daily activities

The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so how much?

***1 - yes, limited a lot***
***2 - yes, limited a little***
***3 - no, not limited at all***

1. ① ② ● lifting or carrying groceries
2. ① ② ● climbing several flights of stairs
3. ① ② ● walking several blocks

## 11. Exercise

How many days per week do you engage in aerobic exercise of at least 20 to 30 minutes duration (fitness walking, cycling, jogging, swimming, aerobic dance, active sports)?

⓪ none
① one
● two days
③ three days
④ four days
⑤ five days
⑥ six days
⑦ seven days

## 12. Strength exercises

How many times per week do you do strength building exercises such as sit-ups, pushups, or use weight training equipment?

● none
② once a week
③ twice a week
④ three or more

## 13. Stretching exercises

How many times per week do you do stretching exercises to improve flexibility of your back, neck, shoulders, and legs?

● none
② once a week
③ twice a week
④ three or more

## 14. Breakfast

How often do you eat breakfast, more than just a roll and a cup of coffee?

● eat breakfast every day
② eat breakfast most mornings
③ eat breakfast two to three times per week
④ seldom or never eat breakfast

## 15. Snacks

How often do you eat snack foods between meals (chips, pastries, soft drinks, candy, ice cream, cookies)?

① three or more times per day
② once or twice per day
● few times per week
④ seldom or never eat typical snacks

## 16. Salt

How often do you add salt to your food or eat salty foods (chips, pickles, soy sauce)?

① seldom or never
② some meals
● most meals
④ nearly every meal

**17. Fat intake** Indicate the kinds of foods you usually eat.

**High fat examples:** hamburgers, hot dogs, bologna, steaks, sour cream, cheese, whole milk, eggs, butter, cake, pastry, ice cream, chocolate, fried foods, and many fast foods

**Low fat examples:** lean meats, skinless poultry, fish, skim milk, low fat dairy products, fruit desserts, gelatin, vegetables, pasta, legumes (peas & beans)

① nearly always eat the high fat foods
② eat mostly the high fat foods, some low fat
③ eat both about the same
● eat mostly low fat foods, some high fat
⑤ eat only low fat foods

**18. Breads and grains** Indicate the kinds of breads and grains you usually eat.

**Refined grain examples:** white bread, rolls, regular pancakes and waffles, white rice, typical breakfast cereals, typical baked goods

**Whole grain examples:** whole grain breads, brown rice, oatmeal, whole grain or high fiber cereals

① nearly always eat refined grain products
② eat mostly refined grain products
③ eat both about the same
● eat primarily whole grain products
⑤ eat only whole grain products

**19. Fruits and vegetables** How many <u>servings</u> of fruits and vegetables do you eat <u>daily</u>?

**A serving is:** 1 cup fresh, 1/2 cup cooked, 1 medium size fruit, or 3/4 cup juice

① 1 or less
② two daily
● three daily
④ four daily
⑤ five or more

**20. Number of drinks** How many alcoholic drinks do you usually have **<u>per week</u>**?

**One drink is:** a 12 oz. beer, 5 oz. wine, or 1.5 oz liquor

● seldom or never
② one to seven
③ eight to fourteen
④ fifteen to twenty
⑤ twenty-one or more

**21. Medications** How often do you use drugs or medicines (include prescription and nonprescription) that affect your mood, help you relax, or help you sleep?

① frequently
② sometimes
③ rarely
● never

**22. Smoking status** Mark the appropriate response.

● have never smoked
② quit smoking two or more years ago
③ quit smoking less than two years ago
④ smoke pipe or cigar only
⑤ currently smoke less than ten cigarettes daily
⑥ currently smoke ten or more cigarettes daily

**23. Chewing tobacco** Do you use chewing tobacco?

① yes
● no

**24. Coping status** How well do you feel you are coping with your current stress load?

● coping very well
② coping fairly well
③ have trouble coping at times
④ often have trouble coping
⑤ feel unable to cope any more

**25. Stress signals** Mark any item below that applies to you.

1. ① Minor problems throw me for a loop.
2. ① I find it difficult to get along with people I used to enjoy.
3. ① Nothing seems to give me pleasure anymore.
4. ① I am unable to stop thinking about my problems.
5. ① I feel frustrated, impatient, or angry much of the time.
6. ① I feel tense or anxious much of the time.

**26. Feelings** The next questions are about how you feel things have been with you during the <u>past four weeks</u>. For each question, please give the one answer that comes the closest to the way you have been feeling. How much of the time during the <u>past four weeks</u>...

*1 - all the time*
*2 - most of the time*
*3 - a good bit of the time*
*4 - some of the time*
*5 - a little of the time*
*6 - none of the time*

1. ①②●④⑤⑥ Have you felt calm and peaceful?
2. ①●③④⑤⑥ Did you have a lot of energy?
3. ①②③④●⑥ Have you felt downhearted/blue?
4. ●②③④⑤⑥ Have you been a happy person?
5. ①②③④●⑥ Have you felt worthless, inadequate, or unimportant?
6. ①②③④●⑥ Did you take the time to relax and have fun daily?

**27. Sleep** How often do you get 7 to 8 hours of sleep?

① always
● most of the time
③ less than half the time
④ seldom or never

**28. Job satisfaction** Indicate level of satisfaction.

① very satisfied
● mostly satisfied
③ not very satisfied
④ dissatisfied
⑤ not applicable

**29. Social support** Do you have friends/family with whom you can share problems/get help if needed?

● yes
② no

**30. Seat belts** How often do you wear a seat belt?

● always
② most of the time
③ less than half the time
④ seldom or never

**31. Sun protection** Do you use sun screen, wear protective clothing, avoid sun bathing, etc.?

● yes
② no

**32. Lifting** When lifting heavy objects, how often do you lift with your legs not with your back?

① always ③ less than half the time
● most of the time ④ seldom or don't know

**33. Drinking and driving** Do you sometimes drive when perhaps you've had too much to drink, or ride with such a person?

① yes ● no

**34. Physical exam** When was your last physical examination? Within the last:...

● year ③ three years ⑤ five or more
② two years ④ four years

**35. Women's health issues** Mark all that apply. Men skip to next question.

1. ① Currently pregnant.
2. ① Had PAP smear, within last 1-3 years.
3. ① Had mammogram within last 1-2 years.
4. ① Gave birth before reaching age 30.
5. ① Passed or reached menopause.
6. ① Taking estrogen, female hormones.
7. ① Practice monthly breast self-exam.

**36. Sick days** How many days did you miss from work (or from school if a student) due to illness or injury during the last 12 months?

⓪ 0 ● 2 ④ 4 ⑥ 6 ⑧ 8 ⑩ 10+
① 1 ③ 3 ⑤ 5 ⑦ 7 ⑨ 9

**37. Preventive exams** Mark the preventive exams you have had during the time frame listed:

1. ● bowel exam, or flexible sigmoidoscopy within last 3-10 yrs.
2. ● dental exam, within last year
3. ● flu shot, within last year

**38. Readiness to change** Indicate how ready you are to make the changes or improvements in your health in the following areas:

*1 - no present interest in making a change*
*2 - plan a change in the next 6 months*
*3 - plan to change this month*
*4 - recently started doing this*
*5 - already do this regularly (6 mos. +)*

1. ① ● ③ ④ ⑤ be physically active
2. ① ② ③ ④ ● practice good eating habits
3. ① ② ③ ④ ● avoid smoking or using tobacco
4. ① ② ③ ④ ● lose weight, or maintain healthy weight
5. ① ② ③ ④ ● handle stress well
6. ① ② ③ ④ ● avoid alcohol or drink in moderation
7. ① ② ③ ④ ● live an overall healthy lifestyle

**39. Health interests** Mark any of the following health improvement opportunities that you would like to be personally notified about if available.

| | | | | | |
|---|---|---|---|---|---|
| 1. | ① | Quitting smoking | 12. | ① | Alcohol/drugs |
| 2. | ● | Weight management | 13. | ● | Healthy back |
| 3. | ① | Aerobics to music | 14. | ① | Medical self-care |
| 4. | ① | A walking group | 15. | ① | Stress management |
| 5. | ① | A jogging group | 16. | ① | CPR training |
| 6. | ① | A fitness evaluation | 17. | ① | First aid |
| 7. | ① | Nutrition improvement | 18. | ① | Health evaluation |
| 8. | ① | Cholesterol reduction | 19. | ① | Women's health |
| 9. | ① | Blood pressure control | 20. | ① | Diabetes education |
| 10. | ① | Reducing coronary risk | 21. | ① | Communication skills |
| 11. | ① | Cancer risk reduction | 22. | ① | AIDS/preventing STDs |
| | ① | Do not notify me of health promotion opportunities | | | |

## Optional

Complete only if instructed to do so.

| | Y N | | Y N | | Y N |
|---|---|---|---|---|---|
| 1. | ①②③④⑤ | 5. | ①②③④⑤ | 9. | ①②③④⑤ |
| 2. | ①②③④⑤ | 6. | ①②③④⑤ | 10. | ①②③④⑤ |
| 3. | ①②③④⑤ | 7. | ①②③④⑤ | 11. | ①②③④⑤ |
| 4. | ①②③④⑤ | 8. | ①②③④⑤ | 12. | ①②③④⑤ |

## CLINICAL DATA Staff Use Only

○ mmoles use decimals, otherwise ignore decimals

| Blood Pressure | | | | | | Cholesterol | | | | | | Glucose | | | | Triglycerides | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Systolic | | | Diastolic | | | Total | | | HDL | | | ○ nonfasting | | | | | | |
| 1 | 2 | 6 | | 8 | 0 | | | | | | | | | | | | | |
| ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ |
| ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① | ① |
| ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② | ② |
| ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ |
| ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ |
| ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ |
| ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ |
| ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ |
| ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ |
| ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ |

| Waist Girth | | | Hip Girth | | | Body Composition Test | Sum of skinfolds | | | Known % fat | | use decimal HbA1c | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ① 3-site UML | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ | ⓪ |
| ① | ① | ① | ① | ① | ① | | ① | ① | ① | ① | ① | ① | ① | ① |
| ② | ② | ② | ② | ② | ② | ② 3-site UMM | ② | ② | ② | ② | ② | ② | ② | ② |
| ③ | ③ | ③ | ③ | ③ | ③ | | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ③ |
| ④ | ④ | ④ | ④ | ④ | ④ | ③ 7-site | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ④ |
| ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ |
| ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ④ known % fat | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ |
| ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ | ⑦ |
| ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ | ⑧ |
| ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ |

Product # APC-328

PLEASE DO NOT WRITE IN THIS AREA

2132776

## Figure A2. Physician Evaluation Questionnaire

*Please answer the following questions for each line in the data. Please note that question 4 will only apply to instances where the employee (identified by the "employee id" variable) is a repeat participant (also specified by participation_id). The columns "Survey_q1, Survey_q1_response, Survey_q2, Survey_q3, and Survey_q4" have been created for answers to each of the questions.*

1. Do you believe this employee likely has one or more medical conditions? If yes, list any such conditions. *Note: Just a best guess for the basic medical condition; doesn't have to be definitive. Please put a "yes" or a "1" in Survey_q1 for each observation if they have one or more medical conditions.*

2. Considering each employee's medical condition(s) and state of general health as determined by laboratory and survey data, how seriously ill is the employee? (1=not very sick for the condition(s); 5=very sick for the condition(s)) *Note: Given the data, this is your professional opinion to how sick the patient is. This may differ slightly from the total score resulting from the scoring system depending on how you weight the profile as a whole.*

3. How much do you think each employee's general state of health would impact his/her ability to carry out an 8-hour job involving manual labor? (1=very little; 5=considerably) *Note: This question is focused on how employee health will impact short-term productivity. This is more specific than question 2, as question 2 more broadly focuses on general wellness (some of which may not immediately impact productivity).*

4. How much did the employee's health improve from the last set of tests? (1=worsened considerably, 2=worsened a little, 3=the same, 4=improved a little, 5=improved considerably) Note*: Again, given the data, this is your professional opinion. This may coincide or differ from the change in the total score from the scoring system, depending on how you weight the improvements in the profile as a whole.*

## Figure A3. Collective Physician Scoring System

**Survey Questions: (note that these variables reference the survey data from WellSource)**

-Question 4: +1 pts for each medical problem and taking medication. +0 pts if not taking medication. If a person is not taking a medication for his/her medical problems, we think it's safe to assume the medical problem is not affecting their productivity or ability to complete routine daily tasks

-Question 5: +1 pts for each symptom

-Question 6: +1 pts for moderate pain, +2 pts for severe pain, +3 pts for very severe pain. +0 pts for none, very mild or mild pain.

-Question 7: +1 pts for a little bit of limitation, +2 pts for some limitation, +3 pts for quite a bit of limitation, +4 pts for could not do daily work, +0 pts for none at all

-Question 8: +1pts for quite a bit of limitation, +2 pts for extremely limited, +0 pts for none at all, slightly or moderately limited

-Question 9: +1 pts for quite a bit of limitation, +2 pts for extremely limited, +0 pts for none at all, slightly or moderately limited

-Question 10: +2 pts for each activity limited a lot, +1 pts for each activity limited a little

-Question 11: +1 pts for patients who exercise <5 days per week

-Question 20: +1 for 15-20 drinks/week, +2 for 21 or more drinks/week

-Question 22: +1 pts for quit smoking <2 yrs, +1 cigar/pipe, +1 smoke <10 cigs/day, +2 smoke >10 cigs/day

-Question 23: +1 pts for chewing tobacco

**Labs: (note that these variables reference the blood data from LabCorp)**

Hemoglobin A1c: +1 pts for Hemoglobin A1c of 6.5-7.0. +2 pts for Hemoglobin A1c>7.0

Albumin: +1 pts for albumin <3.6

TSH (thyroid): +1 pts for TSH>4.5 or <.45

Hemoglobin: +1 pt for hemoglobin<12.5, +2 pts for hemoglobin<10, +3 pts for hemoglobin<8

C Reactive Protein: +1 pts for CRP>3

Total Cholesterol: +1 pts for total cholesterol>250, +2 for total cholesterol >500

HDL: +1 pts for HDL cholesterol<35

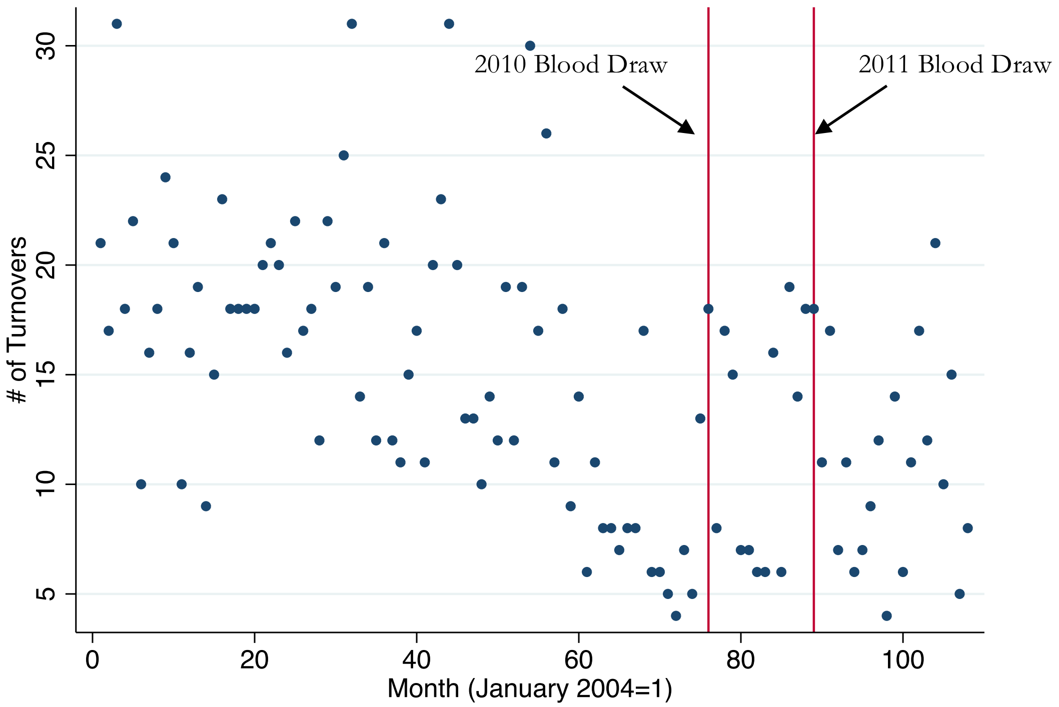
LDL: +1 pts for LDL cholesterol>190

BMI>40: +2 pts, BMI>30: +1 pts, BMI<30: 0 pts

Blood Pressure>140/90: +1 pts

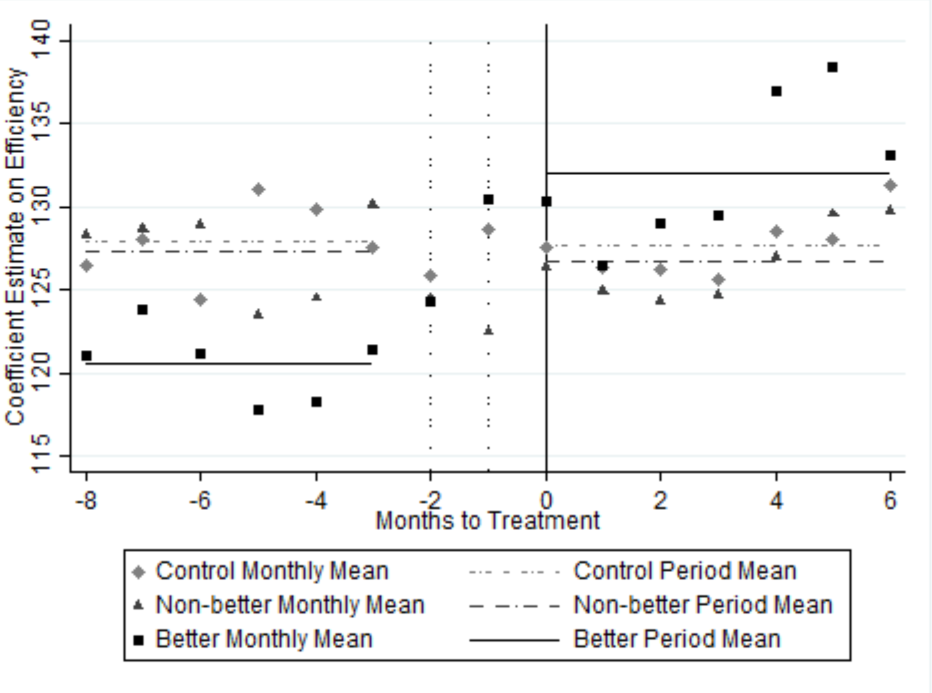
*Lowest Possible Score: 0 pts*
*Highest Possible Score: 50 pts*

## Figure A4: Number of Production Employee Turnovers by Month

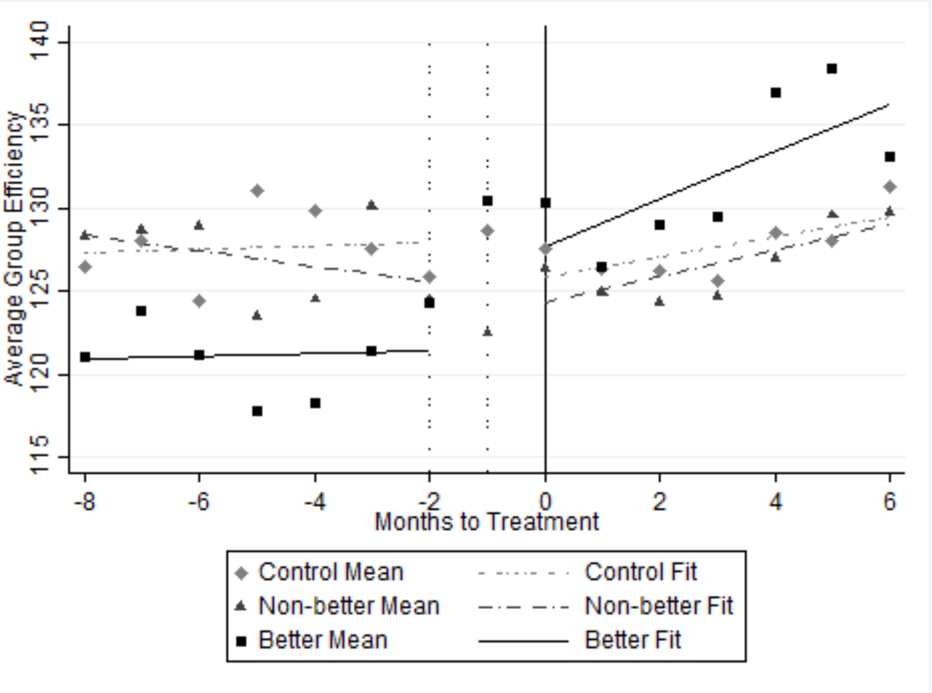


## Figure A5: Raw Data Plots

**A. No trend line – pre- and post-treatment average only**

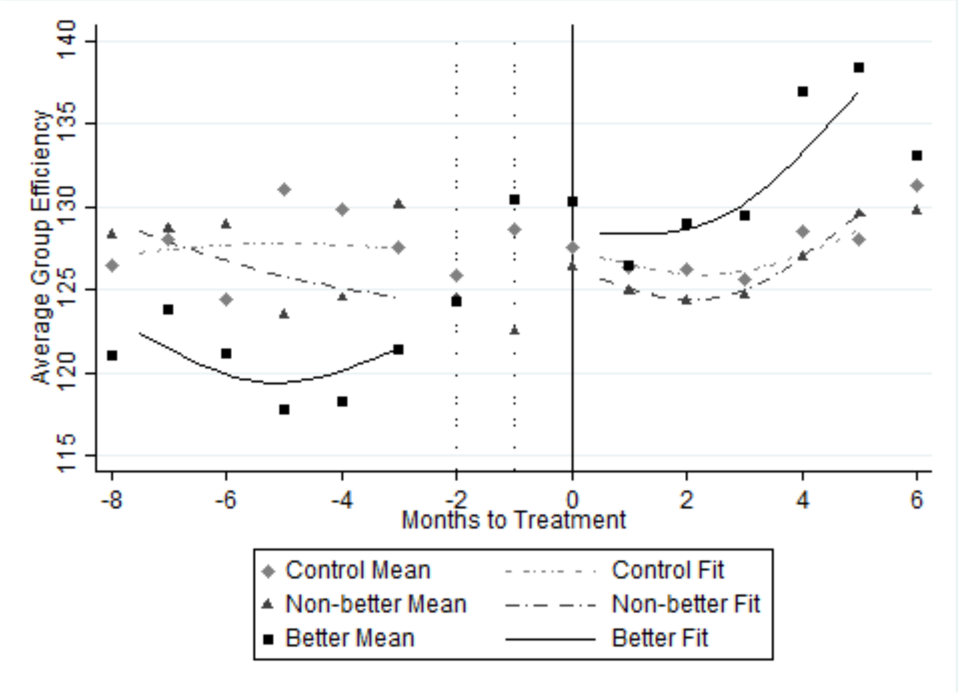


**B. Linear trend line**



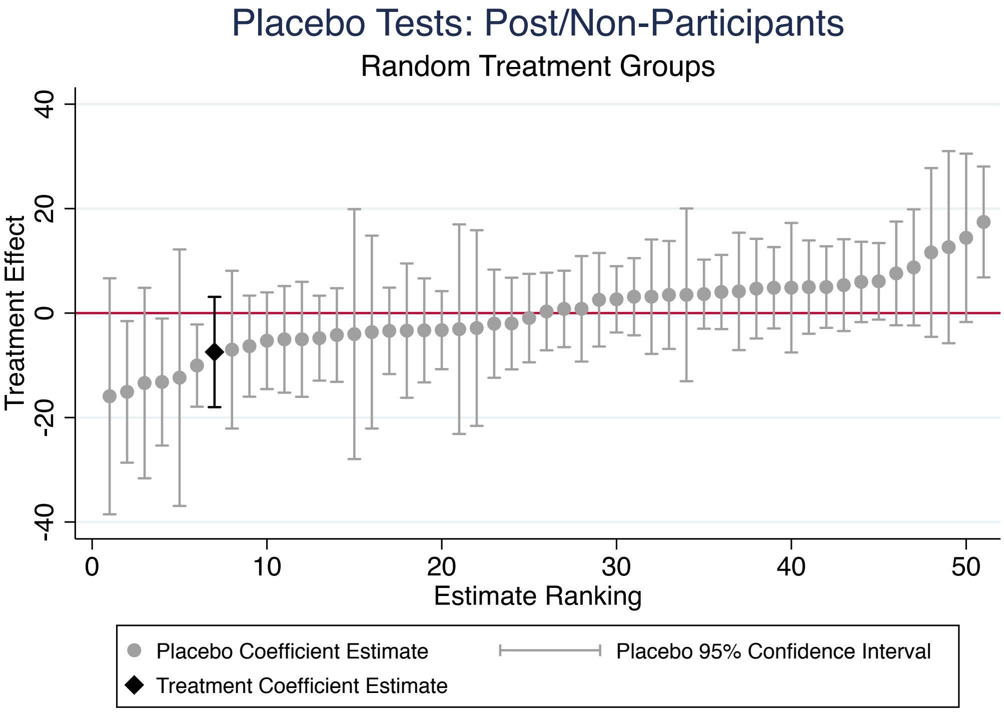
Note: charts show average monthly efficiency by group. Month t=-2 is the date of program announcement. Month t=-1 is the date of the blood draw. Month t=0 is the date of the seminar. Best fit lines are calculated up to t=-2 for the pre-period, and starting at t=0 for the post-period.

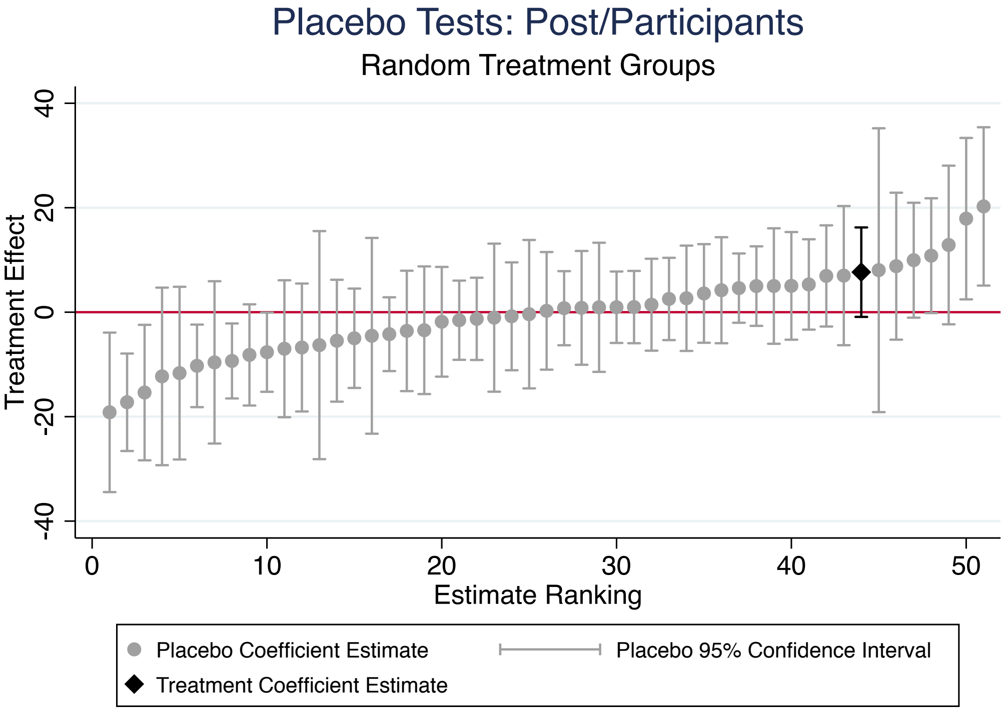
### C. Polynomial spline

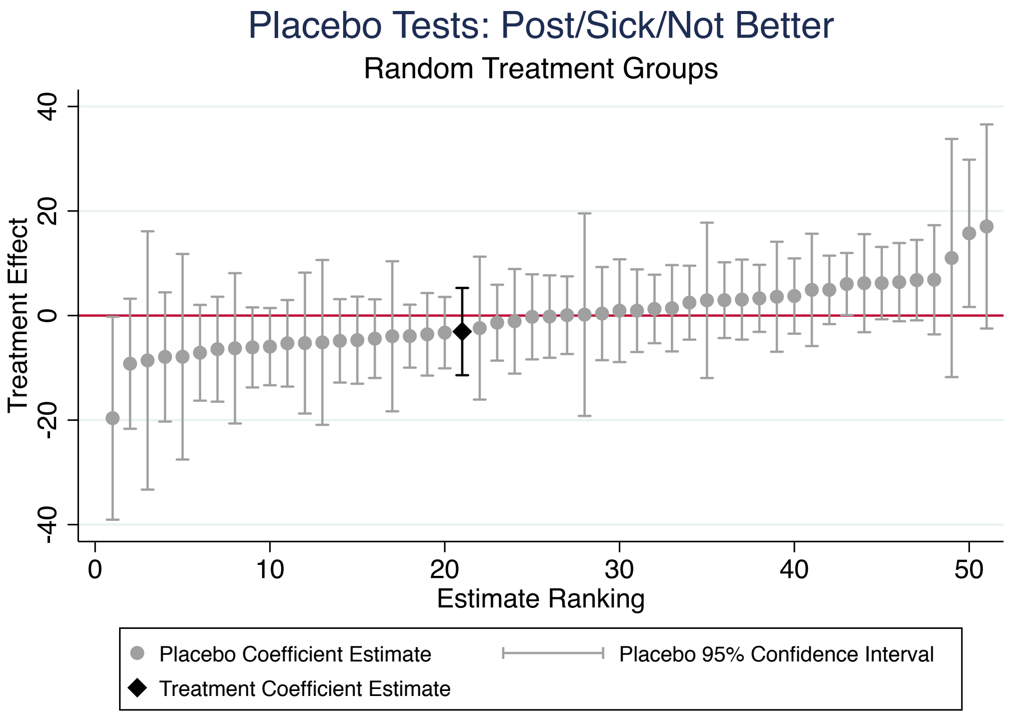


Note: chart show average monthly efficiency by group. Month t=-2 is the date of program announcement. Month t=-1 is the date of the blood draw. Month t=0 is the date of the seminar. Best fit lines are calculated up to t=-2 for the pre-period, and starting at t=0 for the post-period.

## Figure A6: Additional Placebo Tests

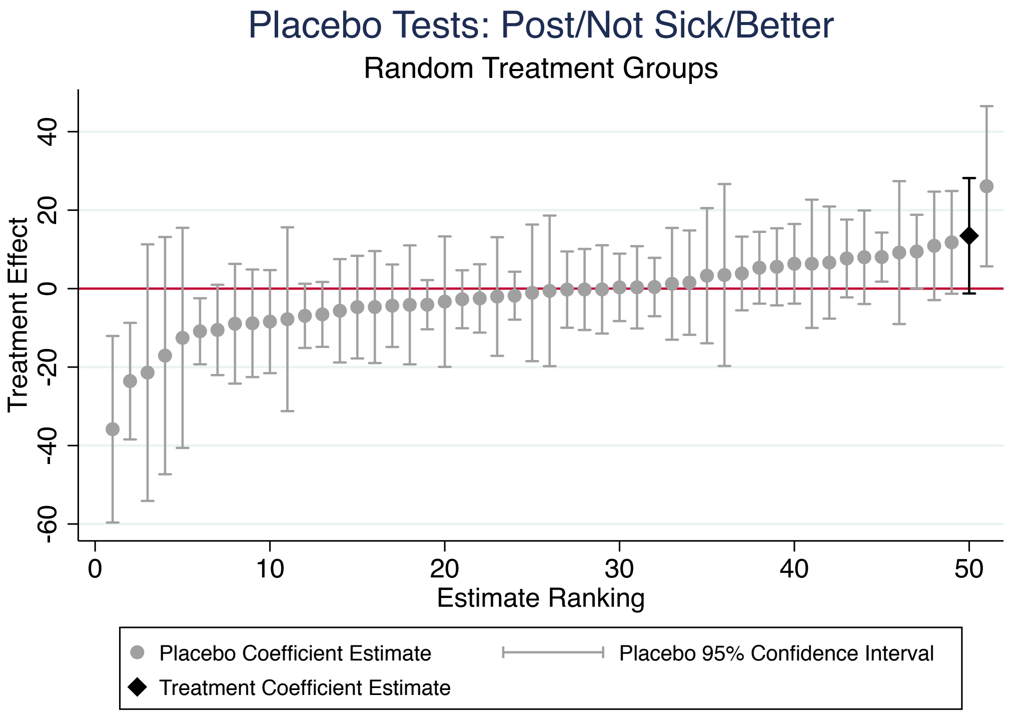






Placebo Tests: Post/Sick/Not Better

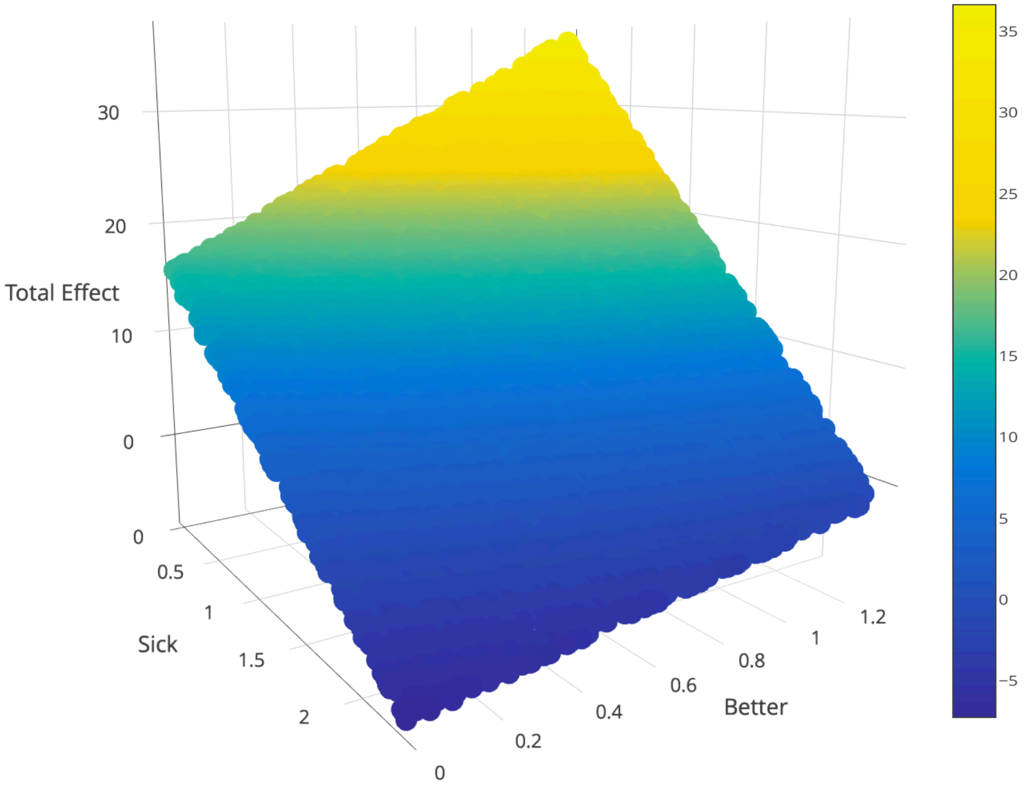
Random Treatment Groups



Placebo Tests: Post/Not Sick/Better

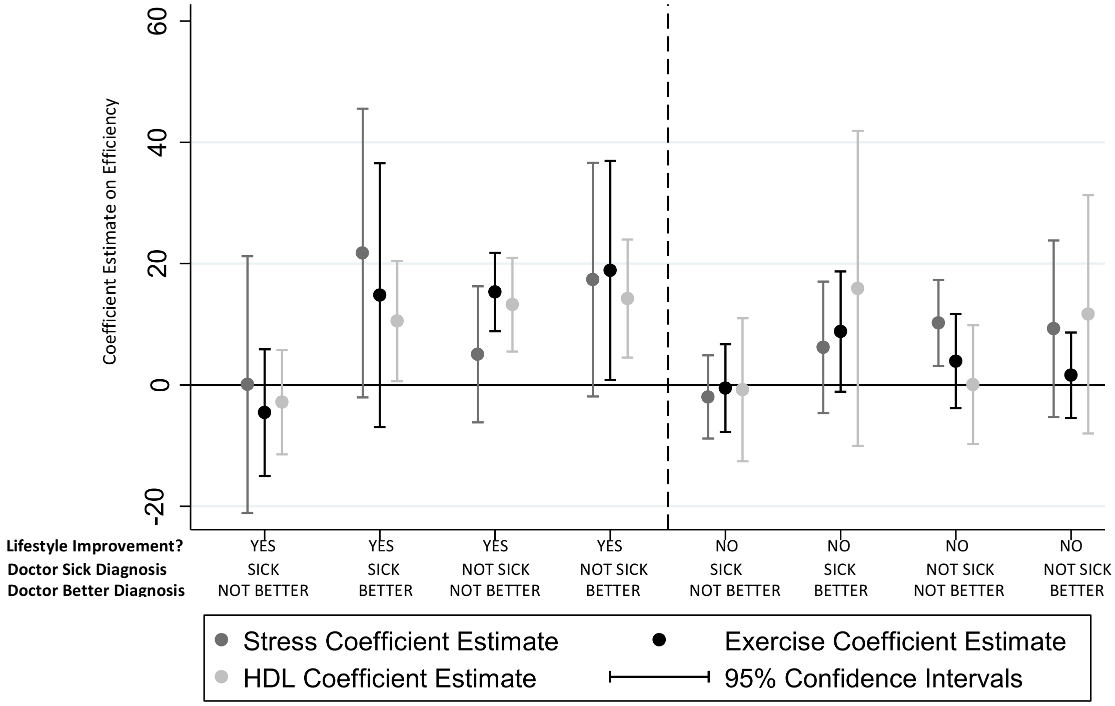
Random Treatment Groups

## Figure A7: Representation of Results Using Continuous Measures



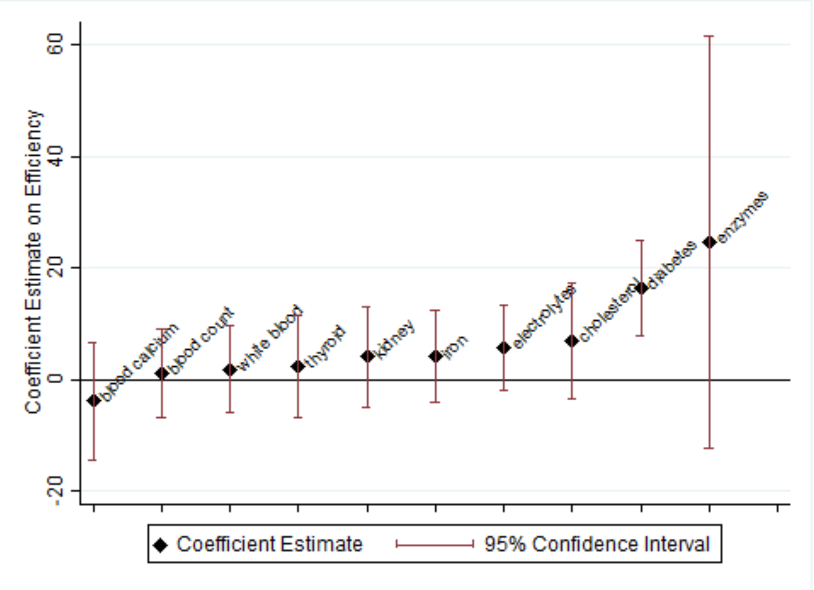
Note: "Sick" is defined from Q2 of the physician survey, and can take a value of 0 (not sick, corresponding to scores of 1, 2 or 3 on the survey), 1 ("somewhat sick," corresponding to a score of 4 on the survey), or 2 ("very sick," corresponding to a score of 5 on the survey). "Better" is taken from Q3 of the physician survey, and can take a value of 0 (not better, corresponding to a score of 1, 2 or 3 on the survey), 1 (a little better, corresponding to a score of 4 on the survey) or 2 (a lot better, corresponding to a score of 5 on the survey.).

## Figure A8: Lifestyle Changes and Efficiency by "*Sick*" and "*Better*"



Note: Effect of lifestyle improvements shown for each possible combination of "sick" and "better" groupings. Lifestyle improvement indicates health improvement on either stress, HDL cholesterol, or exercise. Coefficient estimates represent efficiency regressed on the interaction between each lifestyle improvement, sick, and better.

## Figure A9: Blood Results for Sick Compliers Whose Blood Tests Improve



Note: Diamonds represent coefficient estimates for each specific blood abnormality category following the main specification. Models are estimated using OLS with clustered standard errors at the individual level. “Sick” is defined as having an abnormal blood test in a specific blood category. “Better” is defined as resolving all abnormal blood tests in that specific blood category. Coefficient estimates are plotted for sick individuals that get better (“Better” equal to 1). Estimates based off 2 improvers in blood calcium, 8 in blood count, 5 in white blood, 5 in thyroid, 6 in kidney, 6 in iron, 6 in electrolytes, 7 in cholesterol, 1 in diabetes, and 2 in enzymes.

## Table A1: Demographics and Efficiency for Three Employee Groups, Pre-blood Draw Data

| | Compliers | | Non-Compliers | | Control | |
|---|---|---|---|---|---|---|
| | *mean* | *std. dev* | *mean* | *std. dev* | *mean* | *std. dev* |
| Efficiency | 120.28 | 23.78 | 117.19 | 15.81 | 124.09 | 21.15 |
| Tenure (months) | 48.03 | 65.57 | 8.99 | 9.21 | 96.23 | 92.11 |
| Hours worked | 8.36 | 0.76 | 8.36 | 0.43 | 8.55 | 0.38 |
| Married | 0.34 | 0.48 | 0.29 | 0.49 | 0.14 | 0.35 |
| Base salary | 22387 | 4433 | 21064 | 4006 | 26712 | 2907 |
| Age | 45.40 | 11.22 | 35.89 | 9.87 | 47.45 | 10.66 |
| N | 55 | | 14 | | 42 | |

Note: Pre-blood draw data presented only for those employees who spanned at least two blood draws, which is the sample included in our main regressions.

## Table A2: Results Using Physicians' Collective Scale

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable | Efficiency | Efficiency | Efficiency | Efficiency |
| Model | OLS | OLS | OLS | OLS |
| Physician Scale | Collective Scale | Collective Scale | Collective Scale | Collective Scale |
| Non-compliers | -7.209 | -7.128 | -7.644 | -7.565 |
| | (5.331) [0.179] | (5.244) [0.177] | (5.419) [0.161] | (5.335) [0.159] |
| Compliers | 4.888 | | | |
| | (3.971) [0.229] | | | |
| Non-sick compliers | | 9.392* | | |
| | | (4.818) [0.054] | | |
| Sick compliers | | 0.979 | | |
| | | (4.004) [0.807] | | |
| Non-better compliers | | | 1.715 | |
| | | | (3.901) [0.661] | |
| Better compliers | | | 10.465* | |
| | | | (5.495) [0.059] | |
| Non-sick, non-better compliers | | | | 7.159 |
| | | | | (4.343) [0.102] |
| Sick, non-better compliers | | | | -3.043 |
| | | | | (4.545) [0.505] |
| Non-sick, better compliers | | | | 13.379* |
| | | | | (7.703) [0.085] |
| Sick, better compliers | | | | 7.860 |
| | | | | (5.940) [0.188] |
| Control for worker experience | Y | Y | Y | Y |
| Plant time trends | Y | Y | Y | Y |
| Fixed effects | Y | Y | Y | Y |
| R$^2$ | 0.465 | 0.458 | 0.462 | 0.464 |
| # of employees | 111 | 111 | 111 | 111 |
| Observations | 52293 | 52293 | 52293 | 52293 |
| Sick cutoff | 3 doctors | 3 doctors | 3 doctors | 3 doctors |

Note: Robust standard errors in parentheses, clustered by individual. P-values in brackets. The dependent variable is daily worker efficiency. All three physicians must specify an individual as "sick" for "sick" to take the value of 1. Results are robust to a cutoff of 2 physicians specifying sickness. "Better" uses the collective physician scoring system. Under this system, "Better" is defined as an improved score between the first and second time participating. ** p<.05 * p<0.10

## Table A3: Main Regression Output, Marginal Effects

| | (1) | (2) |
|---|---|---|
| Dependent Variable | Efficiency | Efficiency |
| Model | OLS | OLS |
| Health Assessment | Independent Doctors | Collective Doctor Scale |
| Non-Participant x Post | -7.456 | -7.565 |
| | (5.328) [0.165] | (5.335) [0.159] |
| Participant x Post | 7.662* | 7.159 |
| | (4.321) [0.079] | (4.343) [0.102] |
| Participant x Post x Sick | -10.729*** | -10.202** |
| | (3.893) [0.007] | (4.004) [0.012] |
| Participant x Post x Better | 5.799 | 6.220 |
| | (6.592) [0.880] | (6.802) [0.363] |
| Participant x Post x Sick x Better | 9.046 | 4.683 |
| | (9.745) [0.355] | (9.076) [0.607] |
| Constant | 54.555* | 54.227* |
| | (31.863) [0.090] | (31.868) [0.092] |
| Tenure (months since hire) | Y | Y |
| Plant time trends | Y | Y |
| Fixed effects | Y | Y |
| $R^2$ | 0.465 | 0.464 |
| # of employees | 111 | 111 |
| Observations | 52293 | 52293 |
| Sick cutoff | 3 doctors | 3 doctors |

Note: Robust standard errors in parentheses, clustered by individual. * p<0.10 ** p<0.05 *** p<0.01

## Table A4: Main Results with Block-Bootstrapped Standard Errors

| | (1) | (2) |
|---|---|---|
| Dependent Variable | Efficiency | Efficiency |
| Model | OLS | OLS |
| Health Assessment | Independent Doctors | Collective Doctor Scale |
| Non-compliers | -7.456 | -7.565 |
| | (5.904) [0.207] | (5.913) [0.201] |
| Non-sick, non-better compliers | 7.662* | 7.159 |
| | (4.444) [0.085] | (4.447) [0.107] |
| Sick, non-better compliers | -3.068 | -3.043 |
| | (4.503) [0.496] | (4.850) [0.530] |
| Non-sick, better compliers | 13.461 | 13.379 |
| | (8.918) [0.131] | (9.166) [0.144] |
| Sick, better compliers | 11.778 | 7.860 |
| | (7.703) [0.126] | (6.137) [0.200] |
| | | |
| Tenure (months since hire) | Y | Y |
| Plant time trends | Y | Y |
| Fixed effects | Y | Y |
| $R^2$ | 0.465 | 0.464 |
| # of employees | 111 | 111 |
| Observations | 52293 | 52293 |
| Sick cutoff | 3 doctors | 3 doctors |

Note: Bootstrapped standard errors in parentheses, using 500 repetitions. P-values in brackets. * $p<0.10$ ** $p<0.05$ *** $p<0.01$

## Table A5: Results Dropping Five Non-compliers Absent the Day of Testing

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable | Efficiency | Efficiency | Efficiency | Efficiency |
| Model | OLS | OLS | OLS | OLS |
| Physician Scale | Independent Doctors | Independent Doctors | Independent Doctors | Independent Doctors |
| Non-compliers | -15.477*** | -14.861*** | -16.364*** | -15.704*** |
| | (3.969) [0.000] | (4.169) [0.001] | (4.034) [0.000] | (4.236) [0.000] |
| Compliers | 4.417 | | | |
| | (3.986) [0.270] | | | |
| Non-sick compliers | | 8.545* | | |
| | | (4.858) [0.082] | | |
| Sick compliers | | 0.579 | | |
| | | (3.987) [0.885] | | |
| Non-better compliers | | | -16.364*** | |
| | | | (4.034) [0.000] | |
| Better compliers | | | -16.364*** | |
| | | | (4.034) [0.000] | |
| Non-sick, non-better compliers | | | | 6.898 |
| | | | | (4.355) [0.116] |
| Sick, non-better compliers | | | | -3.319 |
| | | | | (4.255) [0.437] |
| Non-sick, better compliers | | | | 12.827* |
| | | | | (7.393) [0.086] |
| Sick, better compliers | | | | 11.596 |
| | | | | (7.199) [0.110] |
| Control for worker experience | Y | Y | Y | Y |
| Plant time trends | Y | Y | Y | Y |
| Fixed effects | Y | Y | Y | Y |
| $R^2$ | 0.457 | 0.458 | 0.460 | 0.461 |
| # of employees | 106 | 106 | 106 | 106 |
| Observations | 50977 | 50977 | 50977 | 50977 |
| Sick cutoff | 3 doctors | 3 doctors | 3 doctors | 3 doctors |

Note: Robust standard errors in parentheses, clustered by individual. P-values in brackets. The dependent variable is daily worker efficiency. All three physicians must specify an individual as "sick" for "sick" to take the value of 1. Results are robust to a cutoff of 2 physicians specifying sickness. "Better" uses the collective physician scoring system. Under this system, "Better" is defined as an improved score between the first and second time participating.

*** p<.01
** p<.05
* p<0.10

## Table A6: Results Using Continuous Measures of "Sick" and "Better" (Marginal Effects)

|  | (1) |
|---|---|
| Dependent Variable | Efficiency |
| Model | OLS |
| Health Assessment | Independent Doctors |
| Post | 2.204 |
|  | (2.39) [0.358] |
| Non-complier x Post | -7.330 |
|  | 4.32 (0.171) |
| Complier x Post | 16.502* |
|  | (8.554) [0.056] |
| Complier x Post x Sick | -9.427* |
|  | (5.471) [0.088] |
| Complier x Post x Better | 8.576* |
|  | (4.214) [0.071] |
| Complier x Post x Sick x Better | 7.077 |
|  | (4.369) [0.107] |
|  |  |
| Tenure (months since hire) | Y |
| Plant time trends | Y |
| Fixed effects | Y |
| First-stage F statistic | 35.03 |
| # of employees | 111 |
| Observations | 52293 |
| Sick cutoff | 3 doctors |

Note: Robust standard errors in parentheses, clustered by individual. P-values in brackets. The dependent variable is daily worker efficiency. "Sick" is defined from Q2 of the physician survey, and can take a value of 0 (not sick, corresponding to scores of 1, 2 or 3 on the survey), 1 ("somewhat sick," corresponding to a score of 4 on the survey), or 2 ("very sick," corresponding to a score of 5 on the survey). "Better" is taken from Q3 of the physician survey, and can take a value of 0 (not better, corresponding to a score of 1, 2 or 3 on the survey), 1 (a little better, corresponding to a score of 4 on the survey) or 2 (a lot better, corresponding to a score of 5 on the survey.).
* p<0.10

## Table A7: Local Average Treatment Effect Models, Second-Stage Total Effects

| | (1) | (2) |
|---|---|---|
| Dependent Variable | Efficiency | Efficiency |
| Model | OLS | OLS |
| Health Assessment | Independent Doctors | Collective Doctor Scale |
| Post | 2.138 | 2.136 |
| | (2.35) [0.364] | (2.354) [0.364] |
| Complier x Post | 7.662* | 7.159 |
| | (4.321) [0.079] | (4.343) [0.102] |
| Complier x Post x Sick | -5.210 | -5.269 |
| | (5.062) [0.303] | (5.378) [0.327] |
| Complier x Post x Better | 11.137 | 11.025 |
| | (8.214) [0.175] | (8.502) [0.195] |
| Complier x Post x Sick x Better | 9.447 | 5.633 |
| | (7.770) [0.224] | (6.556) [0.390] |
| | | |
| Tenure (months since hire) | Y | Y |
| Plant time trends | Y | Y |
| Fixed effects | Y | Y |
| First-stage F statistic | 35.03 | 35.03 |
| # of employees | 111 | 111 |
| Observations | 52293 | 52293 |
| Sick cutoff | 3 doctors | 3 doctors |

Note: Robust standard errors in parentheses, clustered by individual. P-values in brackets. The dependent variable is daily worker efficiency. Column 1 defines "sick" individuals using each physician's evaluation. All three physicians must specify an individual as "sick" for sick to take the value of 1. Results are robust to a cutoff of 2 physicians specifying sickness. "Better" uses the average of the physician's evaluations on improvement (Q4 in the Physician Evaluation Questionnaire), and takes the value of 1 if the average indicates improvement (>3 on the Q4 5-point scale). For Column 2, "sick" is defined the same as column 1, but "better" uses the collective physician scoring system. Under this system, "Better" is defined as an improved score between the first and second time participating.

* $p<0.10$