Illinois Workplace Wellness Study

Pre-Analysis Plan

Damon Jones, University of Chicago and NBER
David Molitor, University of Illinois at Urbana-Champaign and NBER
Laura Payne, University of Illinois at Urbana-Champaign
Julian Reif, University of Illinois at Urbana-Champaign and NBER

July 11, 2016
Updated August 7, 2016
Updated January 11, 2016
Updated July 10, 2017
Updated August 5, 2017
Updated October 6, 2021
Updated March 28, 2022

1. Introduction

This plan outlines the hypotheses to be tested in the analysis of the impact and behavioral response to the introduction of a comprehensive workplace wellness program at the University of Illinois at Urbana-Champaign (UIUC). This pre-analysis plan has been created prior to the collection of data.

We updated this plan on August 7, 2016. We made three changes. First, we removed a treatment group that used a cluster-based assignment. The purpose of that treatment group was to identify peer effects, but we decided that an alternative approach for estimating peer effects, which is outlined in this updated plan, was preferable. Second, we now allow for two different sets of control variables in our regressions. One consists of only the strata variables. The second consists of variables identified by LASSO. Third, we added additional details about a contingency plan that will be executed in the event of low turnout during biometric screening. This updated was submitted prior to the beginning of the study's intervention.

We updated this plan on January 11, 2016. We made following changes:

a. In Section 5.a we added a specification to our selection analysis, where average selection is first estimated before we estimate selection separately within each treatment group. This simpler specification is nested within our original, more flexible specification. At the time of the update, analysis of the data had not yet begun, except for one comparison of mean amount paid for health care claims in the 18 months prior to the study, among those who completed a screening and those who did not. This preliminary statistic was calculated in response to a last-minute funding opportunity. Importantly, our ultimate test of selection will include a richer set of participation stages: screening, HRA, and wellness

activities. This update to the analysis plan took place prior to any analysis of HRA or wellness activity participation.

b. In Section 5.a, we added lagged health claims data to the set of potential control variables to be chosen within a LASSO framework. This update was made prior to any analysis of participation conditional on treatment group.

c. In Section 5.a, we added a treatment effect regression where treatment groups are pooled and compared to the control group. We also added a treatment effects regression that compares members of each treatment group to the control group, allowing for separate effects. Our original specification only compared members of treatment group C to the control group. This update was made before any analysis of post-treatment outcomes.

d. In Section 5.a, we added two new 2SLS regressions. The first features an indicator for participating in at least one wellness activity. The second features a separate indicator for a fall wellness activity and a spring wellness activity. These are now considered in addition to our original specification, which features a model with the count of wellness activities as the regressor. This update was made before any analysis of post-treatment outcomes.

e. In Section 5.b, heterogeneous effects models were extended to include the above-mentioned additions to our specifications. This update was made before any analysis of post-treatment outcomes.

f. We made a nominal change to section 5.e, where we relabeled data previously referred to as "untruthful" to "inconsistent." This edit did not involve any substantive changes to our experimental intervention or methodology.

We updated this plan on July 10, 2017. We made the following changes:

a. In Section 2.b and Section 2.c, we have expanded our follow-up sample to include all study participants. Incentives for the follow-up survey were changed to a flat $20 gift card for all survey respondents. The incentives for the follow-up biometric screening are now randomized between $0 and $125. This update was made on the same day as the launch of the follow-up study.

b. In Section 2.b we have also added an additional year of the study, including Fall and Spring wellness courses in 2017-2018 and a second follow-up survey and biometric screening in the summer of 2018. This update was done prior to the launch of any of these interventions.

c. In Section 3, our description of the follow-up survey and follow-up biometric screening indicate that all study participants will be invited. This update was made on the day of the launch of the follow-up study.

d. In Section 4, Hypothesis C.1 and Hypothesis C.3, follow-up survey variables were correctly labeled as "G" instead of "A." This update has no bearing on the experimental design.

e. In Section 4, Hypothesis C.2, new survey questions on job satisfaction, presenteeism, productivity, and job search were added to the follow-up survey. This update was done before any data was collected from the follow-up survey.

f. In Section 5.f, we include references for the method of bounding treatment effects in the presence of differential attrition. This update was made before any bounding calculations were conducted.

g. In Section 4.c, we added one additional variable, "Retention (C3)". This update was made after receiving HR data from the university that covered the period up through May 31, 2017.

h. We have clarified that, when examining multiple outcomes within a domain, that we will be adjusting standard errors using the step-down procedure of Westfall and Young (1993). We have also removed statements that we will be employing standardized treatment effects when analyzing online survey or HR variables. This update was made after receiving HR data from the university that covered the period up through May 31, 2017, but before any data was collected from the follow-up survey.

We updated this plan on August 5, 2017. We made the following changes:

a. In Section 4, Hypotheses C.1, C.2, and C.3 we added additional outcomes to each subgroup, comprised of indices of all subgroup variables. This includes either a standardized index (following e.g. Finkelstein et al. (2012) or Liebman and Luttmer (2015)) or a subset of the principal components. These changes were made before any follow-up survey data was analyzed, and before follow-up biometric data had been collected.

b. In Section 4, Hypothesis C.2, we added an additional survey response regarding worker's perception of the workplace (G62) and measuring hours worked (G45). These changes were made before any follow-up survey data was analyzed.

We updated this plan on October 6, 2021. We made the following changes:

a. In Section 4, Hypothesis E.2 was added indicating that we will estimate heterogeneity using Bayesian Causal Forests. This change was made before this heterogeneity analysis was initiated.

b. In Section 5.b, we added text describing the methods we will use for the Bayesian Causal Forest (BCF) analysis of heterogeneity.

We updated this plan on March 28, 2022. We made the following changes:

a. In Section 3, we added a description for a new data source I: online survey of predictions.

b. In Section 4, Hypothesis Group F was added indicating that we will compare estimates of the persistence of financial incentives to predictions elicited using the online survey of predictions.

2. Overview of the Study

a. Motivation

Workplace wellness programs have become a $6 billion industry and are widely touted as a way to improve employee well-being, reduce health care costs by promoting prevention, and increase workplace productivity. Yet, there is little rigorous evidence available to support these claims, partly because the voluntary nature of these programs means that participants may differ from nonparticipants for reasons unrelated to the causal effects of the wellness program. We will implement a randomized control trial to identify the effects of incentives on wellness program participation, produce causal estimates of the effect of wellness programs on health outcomes, determine what kinds of employees benefit from wellness programs the most, and test for the presence of peer effects in wellness participation.

b. Experimental Design

Our experiment consists of a baseline survey, followed by random assignment to a control or one of three treatment groups, A, B, or C. Individuals assigned to a treatment group are offered incentives to complete a biometric screening and health risk assessment. Thereafter, members of the treatment group are further given incentives to participate in up to two wellness programs, one in each semester of the school year. One year later, we will follow up with a subset of study participants, again offering a follow-up survey and biometric screening. Survey, biometric screening, and health risk assessment data will be combined with administrative data from health insurers and the university human resources department.

At baseline, employees will be invited via postcard and email to participate in the study by completing an online survey. In return for completing the survey, participants will be given a $30 Amazon.com gift card. Consent will be obtained before the survey is taken.

Following the baseline survey, employees will be assigned to either the control group or a treatment group. Members of the control group have no further intervention in the first year of the program. Members of the treatment group will be invited to participate in the **iThrive** wellness program consisting of the following:

i. First, they are given the opportunity to participate in a biometric screening and health risk assessment. The biometric screening is scheduled on campus and takes approximately 15 minutes. The biometric test will measure: (1) anthropometrics such as height, weight, and waist circumference (to assess obesity and overweight status); (2) resting blood pressure (to assess hypertension); (3) blood glucose (to assess diabetes

risk); and (4) total, LDL, and HDL cholesterol levels, total cholesterol ratio, and triglycerides (to assess risk of cardiovascular disease). Biometric screening is carried out by a third-party vendor, Presence Health.

ii. After completing the biometric screening, the participants are invited to complete an online health risk assessment (HRA). The HRA is questionnaire designed to identify areas of health improvement, by asking a series of questions related to wellness, health status, nutrition, healthy activities, desire to improve health, preventative health measures. The HRA is also prepopulated with biometric information from the above screening. Upon completion, participants are given customized feedback on areas of improvement. The HRA is administered by a third-party vendor, Wellsource.

iii. Upon completion of the HRA, participants are given the option to enroll in up to two wellness courses, one in the fall semester and one in the spring semester. Courses are designed by the UI Wellness Center and include an Active Living class; self-paced online health challenges in physical activity, weight management, and healthy eating; a weight management class; a tobacco cessation hotline; a stress management class; a Tai Chi class; and a chronic disease management class.

Members of the treatment group will be offered financial incentives for completion of the different stages of the iThrive program. Treatment groups A, B and C will receive $0, $100, and $200, respectively, upon successfully completing both the biometric screening and HRA. Within each treatment group, half of the participants will be offered $25 for each wellness course completed, for up to two courses. The other half will be offered $75 for completion of each wellness course. All incentives will be made known to treatment group members at the onset of treatment group assignment. Throughout the first year of the program, members of the treatment group will have access to an online portal that provides information on treatment group assignment, current progress, accrued incentives, scheduling for biometric screening, HRA access, and wellness program enrollment.

One year after the launch of the study, we will administer a follow-up survey and biometric screening among all study participants. Members of the follow-up study will be offered $20 to complete a follow-up survey. We will also randomly offer $0 or $125 for completion of a biometric screening at follow-up, with equal probability. Follow-up study participants will be made aware of these incentives after the first year of the study is complete. Survey incentives will be shared prior to the launch of the follow-up survey and screening incentives will be shared after the follow-up survey, but before the screening.

The timeline of the study is as follows:

i. **July 11 – Aug 1, 2016:** Baseline Survey is administered online
ii. **August 1 – August 7, 2016:** Treatment assignment
iii. **August 8 – September 9, 2016:** Biometric screenings
iv. **August 22 – September 23, 2016:** HRA is administered online
v. **September 26 – December 2, 2016:** Fall wellness courses
vi. **January – May, 2017:** Spring wellness courses
vii. **July – August, 2017**: Follow-up survey
viii. **August – September, 2017**: Follow-up biometric screening
ix. **September – December, 2017:** Fall wellness courses
x. **January – May 2018:** Spring wellness courses
xi. **July – August, 2018:** Second follow-up survey
xii. **August – September, 2018:** Second follow-up biometric screening

c. Sample Selection and Treatment Assignment

The initial pool for our study includes 12,459 benefits eligible employees at the UIUC. Specifically, the set of eligible employees include all University of Illinois employees satisfying three criteria as of June 10, 2016: a) physically located on the UIUC campus, b) not terminated, and c) eligible for benefits through the Illinois Department of Central Management Services. From this set of employees, 15 were excluded due to their direct involvement with the approval, implementation or design of the study—members of the research team, members of the IRB review panel involved in the study design, staff directly involved in collecting program data, and family members of the research team.

Of the 12,459 employees in the initial pool, the "Core Sample" will consist of all employees who respond to a baseline survey. We estimate a response rate of approximately 50%, resulting in a Core Sample of 6,000 employees. All subsample sizes below will be based on this estimate of a 50% baseline survey response rate. We will assign 1,100 employees to treatment group A, 1,100 employees to treatment group B, and 1,100 employees to treatment group C. The remaining 2,700 employees will be assigned to the control group.

In the second year of the study, all study participants will be invited to participate in a follow-up survey and biometric screening.

At the time of taking the baseline survey, participants will be informed that they may be contacted for follow up treatments, but will not know control or treatment group status. After the baseline survey, employees will be assigned to a treatment or control group using pseudo-random numbers generated by a computer program according to the following steps:

i. Employees will be randomly assigned to the control group or treatment group A, B or C. Randomization will be stratified by age, gender, annual salary, race, and employment class (Faculty, Academic Professional, and Civil Service). We will require there to be at least 20 people per strata (two for each control and/or treatment group) in each treatment cell. If there are fewer than 20 people, then we will aggregate strata as necessary.

ii. One year following the original study, all study participants will be invited to participate in the follow-up survey and biometric screening. Incentives for the follow-up screening will likewise be randomized using age, gender, annual salary, race, and employment class strata.

3. Data Sources

a. **Baseline Survey (A)**: Self-reported health, workplace, and demographic information will be collected via an online survey with up to 66 questions. Only a subset of questions are answered by participants, based on skip logic.

b. **Health Insurance Claims Data (B):** As a part of consenting to our study, participants will grant us access to health insurance claims data. These include total costs of services, bill amounts, and diagnosis codes. Health insurance data will be collected from 2015 – 2020. Claims data are currently only available for employees in two insurance companies, which together comprise 70% of employees in the initial pool with health benefits through the University.

c. **Human Resources Data (C):** Participants will also grant us access to UIUC human resources data, including absenteeism, turnover, employment unit, department, tenure, salary, age, race, sex, and benefit elections. HR data will be collected from 2015 – 2020.

d. **Biometric Screening (D):** A third-party vendor will measure height, weight, waist circumference, resting blood pressure, total cholesterol, total cholesterol ratio, HDL cholesterol, LDL cholesterol, triglycerides, glucose levels. In addition, a set of questions will be asked prior to the revelation of the results of the biometric screening, to measure expectations of weight, height, cholesterol level, blood pressure, glucose level, and body mass index. A second biometric screening will be administered among a subset of participants during the second year of the study.

e. **Health Risk Assessment (E):** The health risk assessment survey will measure self-reported wellness, health status, nutrition, healthy activities, desire to improve health, preventative health measures. Answers to the adaptive survey are

combined to create customized indices of health risk.

f. **Wellness Course Data (F):** We will collect data on enrollment, participation and completion of wellness courses. Enrollment is recorded using our online registration system. Participation and completion are monitored by wellness instructors. Completion is defined as participation in at least 80% of the weeks that a course is provided.

g. **Follow-Up Survey (G):** Similar to the baseline survey, offered to all study participants.

h. **Follow-Up Biometric Screening (H):** Identical to the first biometric screening, offered to all study participants.

i. **Online survey of predictions (I):** Survey of individuals familiar with experiments and/or wellness programs, designed to elicit predictions about the persistence of the financial incentives used in the Illinois Workplace Wellness Study. An initial set of survey invitation will be sent to about 570 experts who have attended relevant economics conferences. Once that survey has closed, we will also post an open link on social media.

4. Hypotheses

Using our combined data, we will identify the effects of incentives on wellness program participation, produce causal estimates of the effect of wellness programs on health outcomes, determine what kinds of employees are most likely to select into wellness programs, and test for the presence of peer effects in wellness participation. See Appendix 1 for a glossary of data source references (e.g. A1 is question 1 on our baseline survey). Our hypotheses can be summarized in the following 5 groups:

a. **Participation Outcomes:** As we increase the incentives for completion of the biometric screening, HRA, and wellness programs, we expect the level of participation to increase.

b. **Selection Outcomes:** Participants may differ from the average employee, in terms of health, health care utilization, and productivity. The average level of health among wellness participants compared to all baseline survey participants is theoretically ambiguous. Under **average** advantageous selection, participants are healthier, lower cost, and more productive, while under **average** adverse selection, the opposite is true. In addition, the marginal participant may differ from the average participant, causing the composition of participants to differ as incentives for participation are varied. Under **marginal** advantageous selection, higher incentives draw in healthier, lower cost, and more productive

participants, while under **marginal** adverse selection, the opposite is true.

c. **Short-Run and Medium-Run Health and Productivity Outcomes:** Take-up of biometric screening, HRA, and wellness programs may potentially increase health, wellbeing, satisfaction, and productivity measures measured one year following the intervention. Even though health measures may improve, health care utilization may also increase, causing an ambiguous effect on health care costs in the short-run.

d. **Peer Effects:** Having a close workplace friend or colleague who also participates in a biometric screening, HRA, or wellness activity may increase one's own participation in wellness programs.

e. **Heterogeneous Treatment Effects:** The baseline characteristics of participants may cause differential impacts of the incentives on participation and health outcomes. In addition, the extent to which biometric screening information deviates from elicited expectations may cause differential impacts on wellness participation and health outcomes.

Hypothesis Group A: *Incentives for participation in the biometric screening, HRA, and wellness programs will have a positive effect on participation, increasing in the size of incentive.*

The following indicators will comprise the family of outcomes in this domain:

Participation Outcomes:

i. Scheduled a biometric screening (D10)
ii. Completed a biometric screening (D11)
iii. Completed HRA (E1)
iv. Enrolled in a fall wellness course (F1)
v. Completed fall wellness course (F2)
vi. Enrolled in a spring wellness course (F3)
vii. Complete spring wellness course (F4)

Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993).

Hypothesis Group B: *In the case of **average** advantageous (adverse) selection, employees who select into wellness programs have higher (lower) baseline health, lower (higher) health costs, and higher (lower) productivity. Consequently, differences in baseline health and productivity between participants and non-participants will be positive (negative), while the analogous difference in baseline health costs will be (negative) positive. In addition, under **marginal** advantageous (adverse) selection, the*

differences in baseline health and productivity between participants and non-participants will be negatively (positively) correlated with incentive size, while the analogous difference in baseline health costs will be positively (negatively) correlated with incentive size.

Hypothesis B.1: Under **average** advantageous (adverse) selection, the difference in baseline health measures between participants and non-participants will be positive (negative). Under **marginal** advantageous (adverse) selection, the difference in baseline health measures between participants and non-participants will be negatively (positively) correlated with incentive size.

The following variables will be used to measure baseline health:

Baseline Survey:

i. Had at least one previous health screening (Any of A1-A5, A8-A9=”yes”)
ii. Physically active (A11=”More active”)
iii. Trying to be more active (A12=”Yes” or A13=”Yes”)
iv. Smoking status:
    1. Current smoker: A16=”Yes” and A17=”Every day” or “Some days”
    2. Former smoker: A16=”Yes” and A17”=”Not at all”
v. Other tobacco use (A22 and A23 != “Not at all”)
vi. Drinking:
    1. Drinker: A24!=0
    2. Heavy drinker: A25>=4 if female, A25>=5 if male
vii. Has at least one chronic health condition (A27)
viii. Self-reported health (A28)
    1. Health is excellent or very good
    2. Health is not poor
ix. Problems with physical activities or pain (A29-A31)
    1. A29=”Somewhat”, “Quite a lot”, “Could not do physical activities” or A30 = “Some”, “Quite a lot”, “Could not do daily work” or A31=”Mild”, ”Moderate”, “Severe”, “Very severe”
x. Energy (A32=”An extraordinary amount”, “Quite a lot”)
xi. Emotional health (A33=”Moderately”, ”Quite a lot”, “Extremely”)
xii. Overweight status (A39=”Overweight” or A40=”Very overweight”)
xiii. Bad health status (A40=”High”, =”Very high” or A41=”High” or “Very high” or A42=”High” or “Very high”)
xiv. Sedentary job (A53=”None at all” or =”Some, but less than 1 hour”)

These baseline health variables can be divided up into two different domains: primary outcomes of interest, and secondary outcomes of interest. The first domain (“primary outcomes”) will include the following variables: current smoker, has at least one chronic health condition, both self-reported health questions, overweight status, and problems

with physical activities or pain. Any baseline health variables not in the first domain will be in the second domain ("secondary outcomes"). Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993).

Hypothesis B.2: Under **average** advantageous (adverse) selection, the difference in baseline health costs and utilization between participants and non-participants will be negative (positive). Under **marginal** advantageous (adverse) selection, the difference in baseline health costs and utilization between participants and non-participants will be negatively (positively) correlated with incentive size.

The following variables will be used to measure baseline health costs and utilization:

Baseline Survey:

i. Drug utilization (A34>0 or A35>0)
ii. Physician or ER utilization (A36!="None")
iii. Hospital utilization (A37!="None")

Insurance Claims:

iv. Number of claims (B1)
v. Number of bed days (B2)
vi. Allowed amount for claims (B4)
vii. Amount paid by plan for claim (B5)

Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993).

Hypothesis B.3: Under **average** advantageous (adverse) selection, the difference in baseline productivity between participants and non-participants will be positive (negative). Under **marginal** advantageous (adverse) selection, the difference in baseline productivity between participants and non-participants will be positively (negatively) correlated with incentive size.

The following variables will be used to measure baseline productivity:

Baseline Survey:

i. Days missed (A45)

Human Resources Data:

i. Salary (C1)

ii. Absenteeism rate (C2)

Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993).

Hypothesis B.4: Under **marginal** advantageous (adverse) selection, average health measures, as measured by biometric screening results, will be negatively (positively) correlated with incentive size, conditional on participating in the biometric screening.

The following variables will be used to measure health conditional on biometric screening:

Biometric Screening Data:

i. BMI (D1)
ii. Waist circumference (D2)
iii. Resting blood pressure (D3)
iv. Total cholesterol (D4)
v. Total cholesterol ratio (D5)
vi. HDL cholesterol (D6)
vii. LDL cholesterol (D7)
viii. Triglycerides (D8)
ix. Glucose levels (D9)

A standardized health measure impact will be obtained, as described in our methodology section below. In addition, we will report each outcome separately. Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993).

Hypothesis B.5: Under **marginal** advantageous (adverse) selection, average health measures, as measured by HRA screening results, will be negatively (positively) correlated with incentive size, conditional on participating in the HRA.

The following variables will be used to measure health conditional on HRA participation:

HRA Data:

i. Health risk score (E1)
ii. Nutrition risk score (E2)
iii. Healthy activities risk score (E4)
iv. Readiness to change score (E5)
v. Biometrics score (E6)

A standardized health measure impact will be obtained, as described in our methodology section below. In addition, we will report each outcome separately. Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993).

Hypothesis Group C: *Participation in biometric screening, HRA, and wellness activities will result in increases in health status and productivity one year later. Participation will have an ambiguous effect on health care costs and utilization.*

Hypothesis C.1: Follow-up health status will be positively affected by wellness treatment, as measured by the reduced form relationship between follow-up health and treatment incentive sizes for biometric screening/HRA and/or wellness activity participation. The same will hold for the IV estimate of the effect of participation in biometric screening/HRA and/or wellness activity participation on follow-up health, instrumented for by incentive size.

The following variables will be used to measure follow-up health:

Follow-Up Survey:

i. Had at least one previous health screening (Any of G1-G5, G8-G9=”yes”)
ii. Physically active (G11=”More active”)
iii. Trying to be more active (G12=”Yes” or A13=”Yes”)
iv. Smoking status:
    1. Current smoker: G16=”Yes” and G17=”Every day” or “Some days”
    2. Former smoker: G16=”Yes” and G17”=”Not at all”
v. Other tobacco use (G22 and G23 != “Not at all”)
vi. Drinking:
    1. Drinker: G24!=0
    2. Heavy drinker: G25>=4 if female, G25>=5 if male
vii. Has at least one chronic health condition (G27)
viii. Self-reported health (G28)
    1. Health is excellent or very good
    2. Health is not poor
ix. Problems with physical activities or pain (G29-G31)
    1. G29=”Somewhat”, “Quite a lot”, “Could not do physical activities” or G30 = “Some”, “Quite a lot”, “Could not do daily work” or G31=”Mild”, ”Moderate”, “Severe”, “Very severe”
x. Energy (G32=”An extraordinary amount”, “Quite a lot”)
xi. Emotional health (G33=”Moderately”, ”Quite a lot”, “Extremely”)
xii. Overweight status (G39=”Overweight” or G40=”Very overweight”)
xiii. Bad health status (G40=”High”, ”Very high” or G41=”High” or “Very high” or G42=”High” or “Very high”)
xiv. Sedentary job (G53=”None at all”, ”Some, but less than 1 hour”)

xv. Indices
    1. Standardized index of above variables
    2. Principal component(s) of above variables

Follow-Up Biometric Screening:

xvi. BMI (H1)
xvii. Waist circumference (H2)
xviii. Resting blood pressure (H3)
xix. Total cholesterol (H4)
xx. Total cholesterol ratio (H5)
xxi. HDL cholesterol (H6)
xxii. LDL cholesterol (H7)
xxiii. Triglycerides (H8)
xxiv. Glucose levels (H9)
xxv. Deviation in expected biometrics from actual biometrics (D10)
xxvi. Indices
    1. Standardized index of above variables
    2. Principal component(s) of above variables

Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993).

Hypothesis C.2: Follow-up productivity will be positively affected by wellness treatment, as measured by the reduced form relationship between follow-up productivity and treatment incentive sizes for biometric screening/HRA and/or wellness activity participation. The same will hold for the IV estimate of the effect of participation in biometric screening/HRA and/or wellness activity participation on follow-up productivity, instrumented for by incentive size.

The following variables will be used to measure follow-up productivity and job satisfaction:

Follow-Up Survey:

i. Days missed (G46)
ii. Hours Worked (G45)
iii. Job satisfaction
    1. G53=”Very satisfied”
    2. G53=”Very satisfied”, ”Somewhat satisfied”
    3. G54=”Yes”
    4.
iv. Presenteeism
    1. Stanford Presenteeim Scale (SPS-6), using G47-G52

v. Productivity
    1. G56=”Very productive”,”Somewhat productive”
    2. G56=”Yes”
vi. Job Search
    1. G64=”Very likely”
    2. G64=”Very likely”,”Somewhat likely”
vii. Workplace Perceptions
    1. G62 = “Very high priority”,”Some priority”
viii. Indices
    1. Standardized index of above variables
    2. Principal component(s) of above variables

Human Resources Data:

ix. Salary (C1)
x. Absenteeism rate (C2)
xi. Retention (C3)

Finally, the HR data allow us to additionally analyze the dynamics of these outcomes using an event-study framework.

Hypothesis C.3: Follow-up health costs will be decreased by wellness treatment, as measured by the reduced form relationship follow-up health costs and between treatment incentive sizes for biometric screening/HRA and wellness activity participation. The same will hold for the IV estimate of the effect of participation in biometric screening/HRA and wellness activity participation, instrumented by incentive size.

The following variables will be used to measure follow-up health costs and utilization:

Follow-Up Survey:

i. Drug utilization (G34>0 or G35>0)
ii. Physician or ER utilization (G36!=”None”)
iii. Hospital utilization (G37!=”None”)
iv. Indices
    1. Standardized index of above variables
    2. Principal component(s) of above variables

Insurance Claims:

v. Number of claims (B1)
vi. Number of bed days (B2)
vii. Allowed amount for claims (B4)

viii. Amount paid by plan for claim (B5)
ix. Indices
    1. Standardized index of above variables
    2. Principal component(s) of above variables

Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993). Finally, the insurance claims data allow us to additionally analyze the dynamics of these outcomes using an event-study framework.

Hypothesis Group D: *The share of one's peers who are induced to participate in the biometric screening/HRA and/or wellness programs increases one's own likelihood to participate in these activities.*

The following indicators will comprise the family of outcomes in this domain:

Participation Outcomes:

i. Scheduled a biometric screening (D10)
ii. Completed a biometric screening (D11)
iii. Completed HRA (E1)
iv. Enrolled in a fall wellness course (F1)
v. Completed fall wellness course (F2)
vi. Enrolled in a spring wellness course (F3)
vii. Complete spring wellness course (F4)

A standardized participation outcome impact will be obtained, as described in our methodology section below. In addition, we will report each outcome separately. Whenever we analyze these variables together, we will adjust the standard errors using the step-down procedure of Westfall and Young (1993).

In addition, peer networks will be identified based on self-reported close friends at work, via the baseline survey:

Baseline Survey:

i. Close co-workers (A43)

Hypothesis Group E: The effect of treatment incentives on participation and short-run outcomes will vary based on baseline characteristics.

Hypothesis E.1: We will look for heterogeneity in participation with respect to the following:

i. Age (C3 ≥ 50)
ii. Sex (C4)
iii. Race (C5, white v. Non-white)
iv. Ethnicity (C6, Hispanic v. Non-Hispanic)
v. Baseline health status
    1. All categories from baseline survey measuring health (these are outlined above in Hypothesis B.1)
vi. Annual Salary (C1, above median)
vii. Employment Class (C7)
viii. Deviation in expected biometrics from actual biometrics (D10)

Hypothesis E.2: We will estimate heterogeneity using Bayesian Causal Forests (BCF). The outcome variables in this analysis will include all outcomes reported in Jones, Molitor, and Reif (2019). The analysis will fit both a standard BCF model and a BCF model with non-compliance. The analysis will summarize posteriors to determine important covariates driving heterogeneity, as in, for example, Woody, Carvalho, and Murray (2021).

Hypothesis Group F: *Financial incentives have a persistent effect on the biometric screening participation rate, and this effect differs from the effect forecasted by individuals familiar with experiments and/or wellness programs.*

Participation Outcome:

i. Completed a biometric screening

Online survey of predictions:

a. Beliefs (point estimate and 95% credence interval) about the effect of the *first*-year incentives on *second*-year screening participation
b. Beliefs (point estimate and 95% credence interval) about the effect of the *first*-year incentives on *third*-year screening participation
c. Beliefs (point estimate and 95% credence interval) about the effect of the *second*-year incentives on *third*-year screening participation

The survey beliefs (including the uncertainty about those beliefs) will be used to construct a prior probability distribution. Using Bayesian methods, we will combine this prior with the estimates from our experiment to produce a posterior probability distribution. This analysis will be performed separately for people recruited over email and those recruited over social media (see Section 3, data source i), and also on the combined sample.

5. Estimation Methodology

a. Treatment Effects

For our main participation outcomes, we will restrict analysis to members of treatment groups A, B, and C. First, we will pool groups B and C, and compare them to group A, as follows:

$$P_i = \alpha + \beta_{B,C} T_{i,B,C} + \gamma_{75} T_{i,75} + \Gamma X_i + \varepsilon_i$$

where P_i is the participation outcome, $T_{i,B,C}$ is an indicator for membership in treatment group B or C, $T_{i,75}$ is an indicator for receiving the $75 incentive for wellness program completion, and X_i is a vector of baseline control variables. We will consider two sets of variables to include in X_i. The first set will include only variables that were used for stratification (i.e., employee class, gender, age, annual salary, and race). The second set will include only variables that are good predictors of the outcome variable. We will identify those control variables by estimating a LASSO regression with five-fold cross validation. We will include in that LASSO regression all available variables from our baseline survey, HR data, and lagged health claims data. We will present one table that shows how results differ using these two different sets of control variables. We will then choose one set as our preferred specification for the remaining tables.

Next, we will separately estimate the effect of treatments B and C, relative to treatment A, as follows:

$$P_i = \alpha + \beta_B T_{i,B} + \beta_C T_{i,C} + \gamma_{75} T_{i,75} + \Gamma X_i + \varepsilon_i$$

where now $T_{i,B}$ and $T_{i,C}$ are indicators for membership in treatment groups B and C, respectively.

Our next set of estimates will test for adverse or advantageous selection, again among members of treatment groups A, B, and C. First, we will estimate the following regression:

$$X_i = \alpha + \delta P_i + \varepsilon_i,$$

where X_i is a baseline variable measured from either the baseline survey, administrative data, or health insurance data. To test for **average** advantageous or adverse selection, we will test the sign of δ. We will then estimate the following regressions:

$$X_i = \alpha_A T_{i,A} + \alpha_B T_{i,B} + \alpha_C T_{i,C} + \delta_A T_{i,A} \times P_i + \delta_B T_{i,B} \times P_i + \delta_C T_{i,C} \times P_i + \varepsilon_i$$

where $T_{i,A}$ is an indicator for membership in treatment group A. To further test for **average** advantageous or adverse selection, we will test the signs of the δ

coefficients. To test for **marginal** advantageous or adverse selection, we will compare the magnitudes of the δ coefficients across treatment groups. In the case that a variable is top-coded, we will use a censored regression model to account for the structure of the limited dependent variable.

Our second set of selection estimates are estimated conditional on either completing the biometric screening or the HRA. In particular, we will estimate the following:

$$X_i = \alpha + \pi_B T_{i,B} + \pi_C T_{i,C} + \varepsilon_i$$

In this case, we can only test for **marginal** advantageous or adverse selection by comparing the magnitudes of the π coefficients. In the case that a variable is top-coded, we will use a censored regression model to account for the structure of the limited dependent variable.

In the above specifications, we will extend the analysis to test for selection with respect to the wellness program incentives, by adding an interaction term $T_{75} \times P_i$.

Our next set of treatment effects measure the short-run effect of wellness participation on health, health utilization, and productivity. We will compare members of the control group to all members of the treatment group in the following regression:

$$Y_i = \alpha + \theta T_i + \Gamma X_i + \varepsilon_i,$$

where Y_i is one of our above-mentioned outcomes of interest and T_i is an indicator variable for membership in any of the treatment groups. Next, we will estimate a more flexible specification as follows:

$$Y_i = \alpha + \theta_A T_{i,A} + \theta_B T_{i,B} + \theta_C T_{i,C} + \theta_{25} T_{i,25} + \theta_{75} T_{i,75} + \Gamma X_i + \varepsilon_i.$$

Finally, we will use a specification that allows for relatively unrestricted interactions between the screening/HRA and wellness activity incentives:

$$Y_i = \alpha + \theta_{A,25} T_{i,A,25} + \theta_{A,75} T_{i,A,75} + \theta_{B,25} T_{i,B,25} + \theta_{B,75} T_{i,B,75} + \theta_{C,25} T_{i,C,25} + \theta_{C,75} T_{i,C,75} + \Gamma X_i + \varepsilon_i,$$

where $T_{i,j,k}$ is an indicator for membership in treatment group $j \in \{A, B, C\}$ with a wellness activity incentive of $k \in \{25,75\}$.

In addition, we will estimate local average treatment effects of HRA/biometric completion and wellness program completion using a two-stage least squares regression (2SLS):

$$Y_i = \alpha + \gamma_{HRA/Screen} HRA/Screen_i + \gamma_{Wellness} Wellness_i + \Gamma X_i + \varepsilon_i$$

where $HRA/Screen_i$ is an indicator for completion of the screening and $Wellness_i$ is an indicator for completion of at least one wellness activity. We will instrument for these regressors using indicators for the 6 treatment groups. Alternatively, we will estimate the following specification:

$$Y_i = \alpha + \gamma_{HRA/Screen} HRA/Screen_i + \gamma_{Wellness,Fall} Wellness_{i,Fall} + \gamma_{Wellness,Spring} Wellness_{i,Spring} + \Gamma X_i + \varepsilon_i,$$

where $Wellness_{i,Fall}$ and $Wellness_{i,Spring}$ are indicators for finishing either a fall or spring wellness activity. Finally, we will estimate a model with a linear effect of wellness activities:

$$Y_i = \alpha + \gamma_{HRA/Screen} HRA/Screen_i + \gamma_{Wellness} \#Wellness_i + \Gamma X_i + \varepsilon_i,$$

where $\#Wellness_i$ is a count of the number of wellness activities completed.

In certain cases, i.e. when looking at productivity measures from HR data and health cost and utilization measures from insurance claims data, we will have longitudinal data, which allows us to look at the dynamics of outcomes over time. In this case, we will also conduct event-study analysis using the following specification:

$$Y_{it} = \alpha + \sum_{j=-k}^{k} \lambda_j D_{it}^{j} + \eta_i + \phi_t + \Gamma \mathrm{X}_{it} + \varepsilon_{it}$$

where D_{it}^{j} is a dummy variable indicating treatment assignment has initiated j periods ago, and η_i and ϕ_t are individual and time fixed effects, respectively.

In order to test for peer effects, we will estimate how participation in an individual's peer group effects the individual's own participation outcome:

$$P_i = \alpha + \rho \bar{P}_{-i} + \Gamma X_i + \varepsilon_i$$

where $\bar{P}_{-i}$ is the leave-out mean of participation in individual i's peer group. We will address the endogeneity of $\bar{P}_{-i}$ by instrumenting for it using variables indicating which treatment group, if any, the peers were assigned to. In this analysis, standard errors will be clustered by peer group.

In addition to the previous specification, we calculate exact p-values for null hypotheses regarding treatment spillovers using the method of Athey, Eckles,

and Imbens (2015).

b. Heterogeneous Effects

In order to identify heterogeneous effects in participation, we will interact baseline characteristics with our treatment effect estimates. We will estimate the following equation:

$$P_i = \alpha + \beta_{B,C} T_{i,B,C} + \gamma_{75} T_{i,75} + \lambda_{B,C} T_{i,B,C} \times X_i + \lambda_{75} T_{i,75} \times X_i + \Gamma X_i + \varepsilon_i$$

We will also test for heterogeneity under the more flexible specification:

$$\begin{aligned} P_i \quad = \quad & \alpha + \beta_B T_{i,B} + \beta_C T_{i,C} + \gamma_{75} T_{i,75} + \lambda_B T_{i,B} \times X_i \\ & + \lambda_C T_{i,C} \times X_i + \lambda_{75} T_{i,75} \times X_i + \Gamma X_i + \varepsilon_i \end{aligned}$$

Our estimates of heterogeneity in the case of short-run health and productivity effects will be obtained as follows in reduced form:

$$Y_i = \alpha + \theta T_i + \phi T_i \times X_i + \Gamma X_i + \varepsilon_i.$$

We will also estimate the following, more flexible specifications:

$$\begin{aligned} Y_i = \alpha + \theta_A T_{i,A} + \theta_B T_{i,B} + \theta_C T_{i,C} + \theta_{25} T_{i,25} + \theta_{75} T_{i,75} + \phi_A T_{i,A} \times X_i \\ + \phi_B T_{i,B} \times X_i + \phi_C T_{i,C} \times X_i + \phi_{25} T_{i,25} \times X_i + \phi_{75} T_{i,75} \times X_i + \Gamma X_i \\ + \varepsilon_i \end{aligned}$$

and

$$\begin{aligned} Y_i = \alpha + \theta_{A25} T_{i,A,25} + \theta_{A,75} T_{i,A,75} + \theta_{B,25} T_{i,B,25} + \theta_{B,75} T_{i,B,75} + \theta_{C,25} T_{i,C,25} \\ + \theta_{C,75} T_{i,C,75} + \phi_{A25} T_{i,A,25} \times X_i + \phi_{A75} T_{i,A,75} \times X_i + \phi_{B25} T_{i,B,25} \times X_i \\ + \phi_{B75} T_{i,B,75} \times X_i + \phi_{C25} T_{i,C,25} \times X_i + \phi_{C75} T_{i,C,75} \times X_i + \Gamma X_i + \varepsilon_i. \end{aligned}$$

Likewise, we will estimate heterogeneous 2SLS estimates:

$$\begin{aligned} Y_i \quad = \quad & \alpha + \gamma_{HRA/Screen} HRA/Screen_i + \gamma_{Wellness} Wellness_i + \Gamma X_i \\ & + \lambda_{HRA/Screen} HRA/Screen_i \times X_i + \lambda_{Wellness} Wellness_i \times X_i + \varepsilon_i \end{aligned}$$

where $Wellness_i$ will either be an indicator for taking at least one wellness activity, a flexible set of indicators for either a fall or spring wellness activity, or the count of completed wellness activities.

One of our interaction terms, the deviation between expected biometrics screening results and actual screening results, is measured conditional on taking

the biometric screening, and, furthermore, could be influenced by treatment assignment. We will first test for significant differences in this measure across treatment groups, and will only include it if the measure does not systematically vary with treatment group.

Regarding the Bayesian Causal Forest (BCF) analysis of heterogeneity, in addition to standard descriptive statistics about the posterior draws, we will analyze the high-dimensional posterior through posterior summarization, namely CART (classification and regression trees) and GAM (generalized additive models) descriptions of the posterior. This includes standard BCF for estimating heterogeneous effects of the ITT as well as BCF-LATE for estimating heterogeneous effects of the compliers' causal effects when accounting for one-sided compliance (not all who were invited to participate in the wellness program did in fact participate, but those who were not invited were prohibited from participating).

c. Multiple Outcomes, Standardized Treatment Effects, Multiple Inference

Many of our domains of interest, such as "baseline health", can be measured by several different variables. When reporting estimates for individual outcomes within a domain, we will report both regular p-values and p-values adjusted to account for multiple outcomes (Westfall and Young 1993). We will follow the methodology employed by Finkelstein et al. (2012) in all cases.

d. Missing Data and Questions with limited Variation

If the outcome variable is missing for a substantial fraction of the sample, we will not conduct regression analysis using that outcome. If that happens, we will note it in our write-up. Also, we will not report results if the outcome variable is identical for 90% or more of the sample. If a control variable is missing for a significant fraction of the population, we will omit it from our regressions. For control variables that are missing for a minor share of respondents, we will use a dummy variable to indicate a missing value, and retain the control variable. Decisions to include or exclude control variables on this basis will be done prior to treatment assignment and observation of outcome variables.

e. Inconsistent responses

The baseline survey asks respondents to report their age and their gender. These data are also available from our administrative HR dataset, so we will be able to assess whether or not these data are consistent with answers on the baseline survey. We will define a respondent as having "inconsistent data" if they report a different age or gender than what is recorded in HR data. We will report as a robustness test results omitting those participants with inconsistent answers. If

omitting these individuals significantly affects our results, then we will remove them from our preferred specification.

f. Follow-Up Survey Attrition

Our follow-up analysis can only be conducted among employees that remain employed at the university and furthermore, among employees that agree to complete the follow-up survey and biometric screening. We will first examine If treatment assignment is related to attrition at these various levels at a statistically significant level. If not, then we will conduct analysis ignoring attrition. However, if we do find a significant difference in attrition, we will implement bounds on our treatment effect estimates, using "Lee bounds" in the case of differential employment turnover, and using sharper bounds to address differential survey and/or biometric screening response, which rely on variation in follow-up incentives to calculate the bounds. We will implement methods outlined in Behaghel et al. (2009) and Behaghel et al. (2015).

6. Contingency Plans

a. Low Baseline Survey Response

If we receive less than 2,000 responses in the 7 days following the launch of the baseline survey, then we will increase the Amazon.com gift card amount from $30 to $50 to encourage a larger response rate. If that happens, we will also award an extra $20 to participants who already responded in the first 7 days.

b. Low Baseline Screening Rates

If rates of participation in the biometric screening are slightly lower than projected, we may assign individuals from the control group to treatment groups A, B, and C in order to increase statistical power. The number of additional individuals invited will depend on the budget and a projected response rate based on early participation in Groups A, B, and C. This decision will be made prior to observation of any outcome variables from the screening rate portion of the experiment.

If rates of participation in the biometric screening are *much* lower than projected, we may add another treatment arm, Group D, comprised of individuals from the control group. Group D will receive the same treatment as Groups A, B, and C, except that the screening incentive will be $400. The number of individuals invited to Group D will depend on the budget and a projected response rate based on early participation in Groups A, B, and C. Determination of whether to add another treatment arm will be done prior to observation of any outcome variables from the screening rate portion of the experiment.