J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

# J-PAL / IRC Evaluating Social Programs Webinar Series

June 15, 22, 29, and July 13, 2020

8 AM to 10 AM EST | 3PM TO 5PM EEST

# Housekeeping notes

- You are all muted and your cameras are off.
- When we move to breakouts, turn your video and audio on
- We recommend choosing the "side-to-side" option in the "View Options" tool at the top of your screen
- If you're having any technical problems with Zoom, use the Chat feature and we will do our best to help
- Today's lecture slides will be posted to the event page after the session

# J-PAL / IRC Evaluating Social Programs Webinar Series

- June 15: Why Randomize?
- June 22: Ethics of Randomized Evaluations
- June 29: Generalizability
- July 13: Building Effective Academic - NGO Partnerships

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

# Why Randomize?
## Lecture & Case Study

Ben Morse
Senior Research, Education, and Training Manager
J-PAL Global

June 15, 2020

# Session Overview

I. Background

II. Why randomize case study

    I. Non-experimental methods

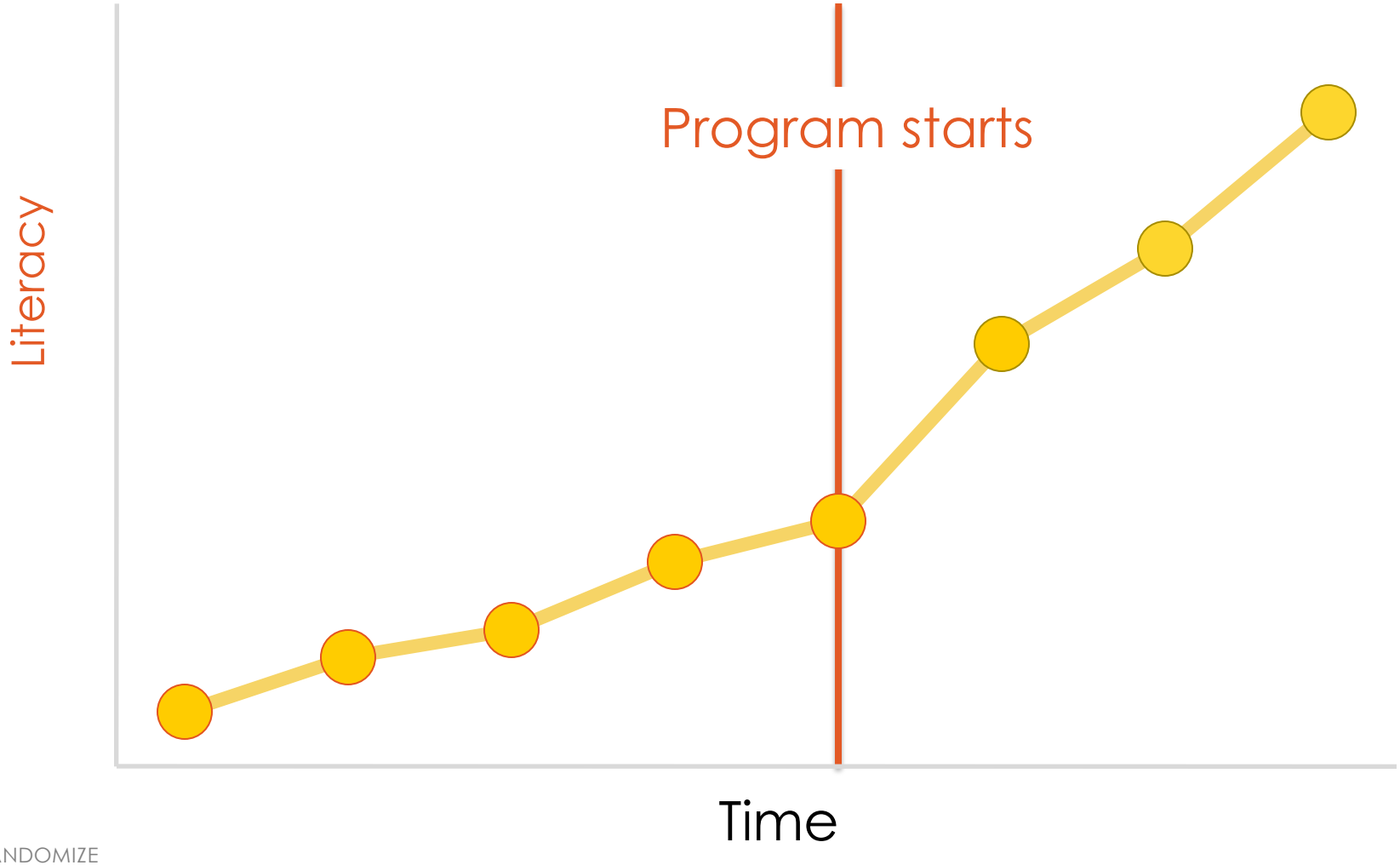
    II. Randomized evaluations

III. Conclusions

# I - BACKGROUND

# What is the impact of Sesame Street on literacy?

# What is the impact of Sesame Street on literacy?



# What is the impact of Sesame Street on literacy?

# How to measure impact?

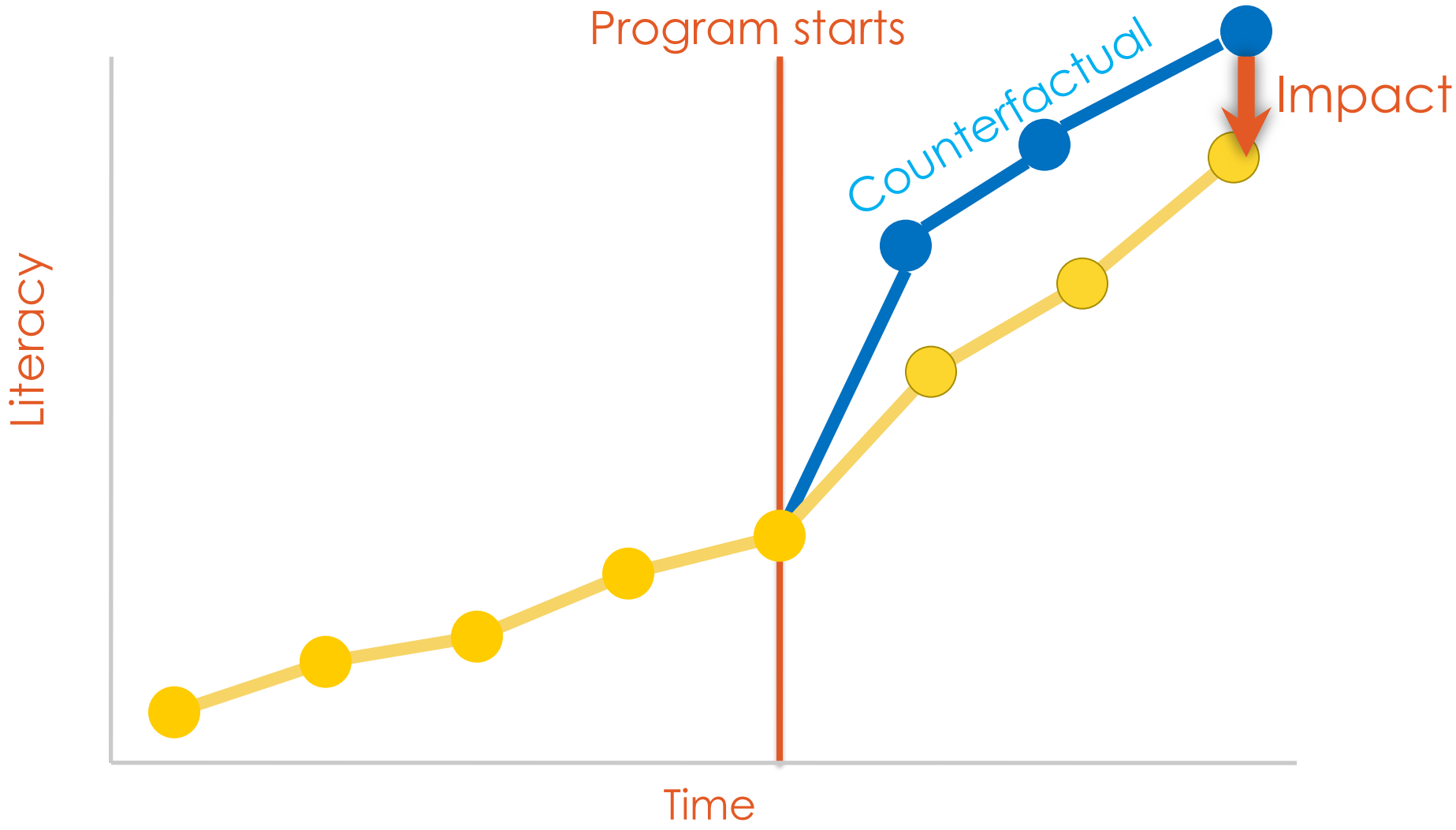
*Impact* is defined as the difference between:

1. the outcome some time after the program has been introduced (the *"factual"*)

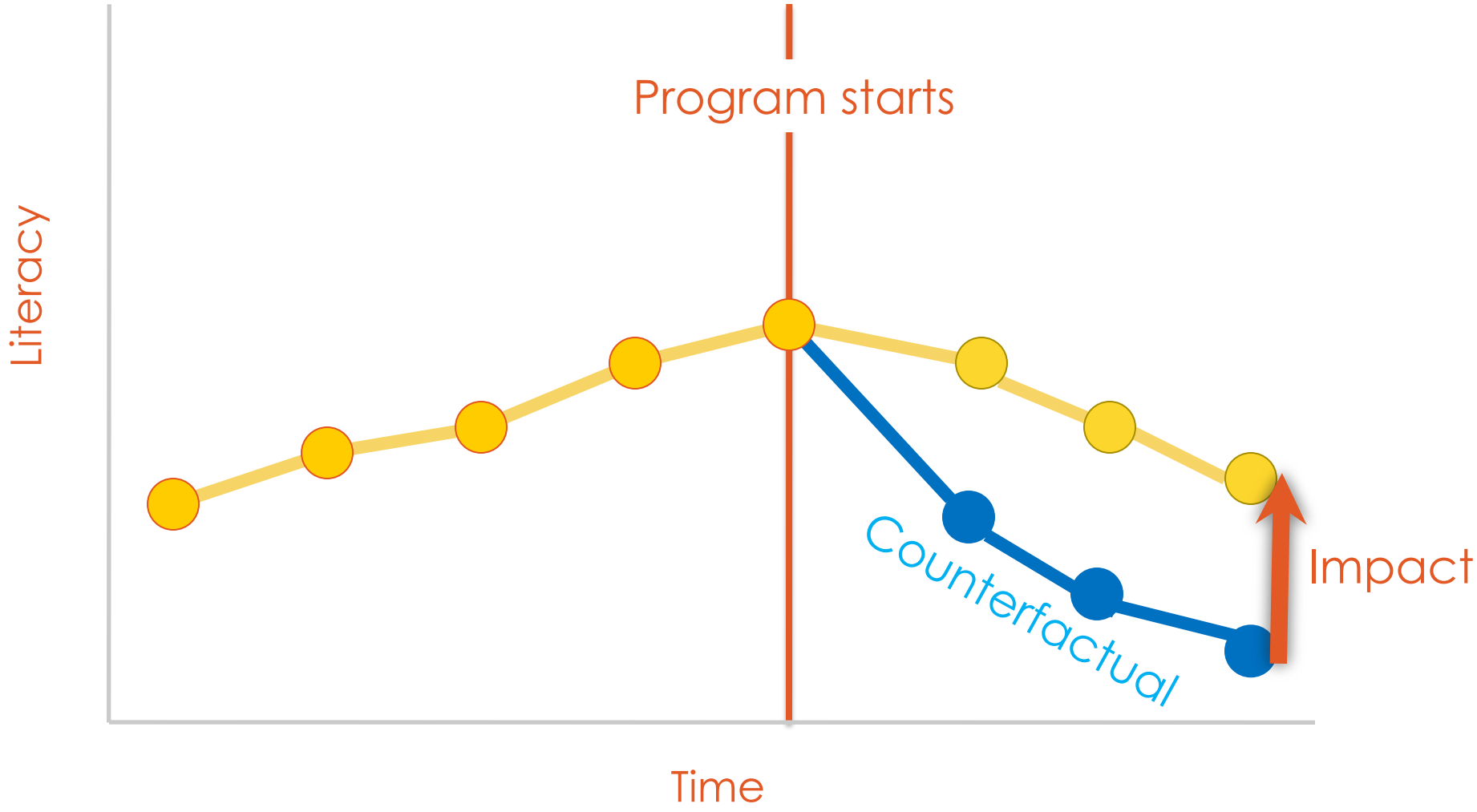
2. the outcome at that same point in time had the program not been introduced (the *"counterfactual"*)

$$factual - counterfactual = impact$$

# Impact: What is it?



# Impact: What is it?



# Counterfactual

The ***counterfactual*** represents the world that program participants would have experienced in the absence of the program

***Problem***: Counterfactual cannot be observed

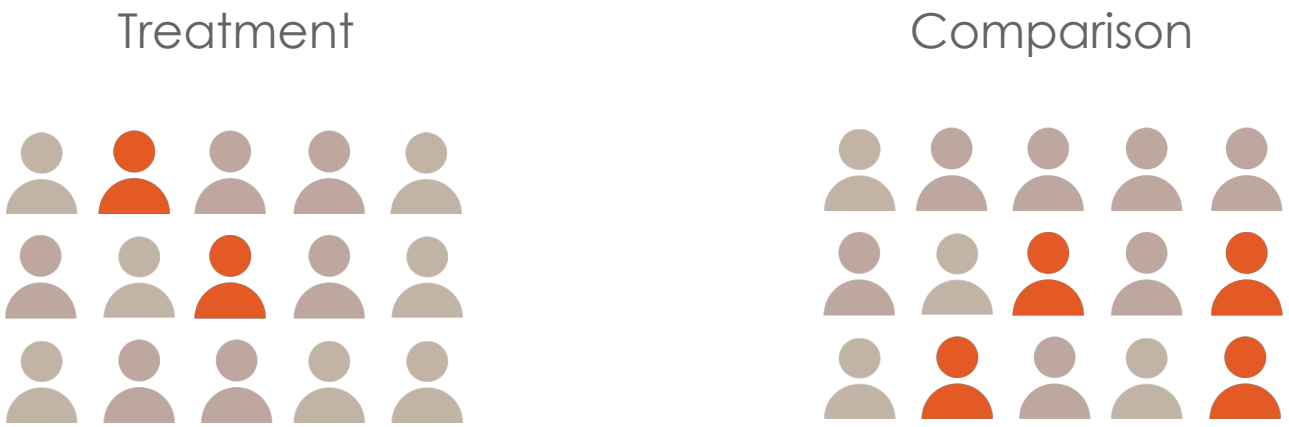
***Solution***: We need to "mimic" or construct the counterfactual

# Constructing the counterfactual

- Usually done by selecting a group of individuals that ***did not*** participate in the program
- This group is usually referred to as the ***control group*** or ***comparison group***
- How this group is selected is a **key decision** in the design of any impact evaluation

# Selecting the comparison group

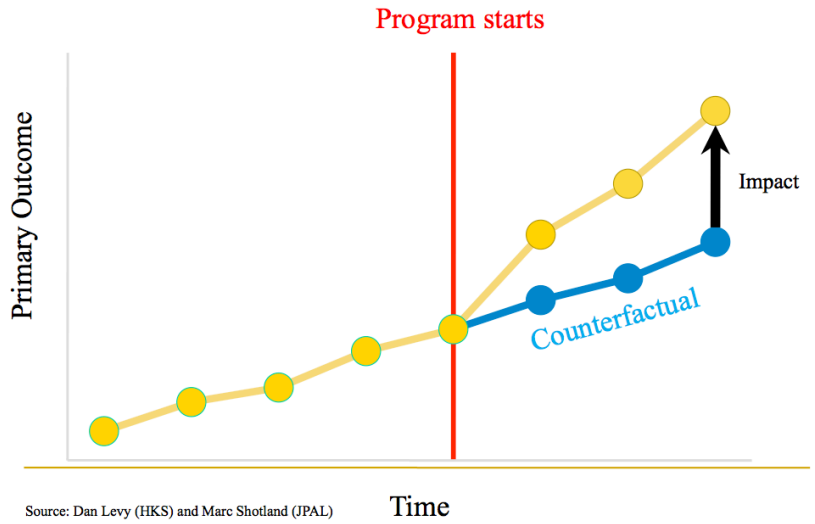
- Idea: Comparability




- Goal: Attribution – any difference between the groups can be attributed to the program, and not to other factors

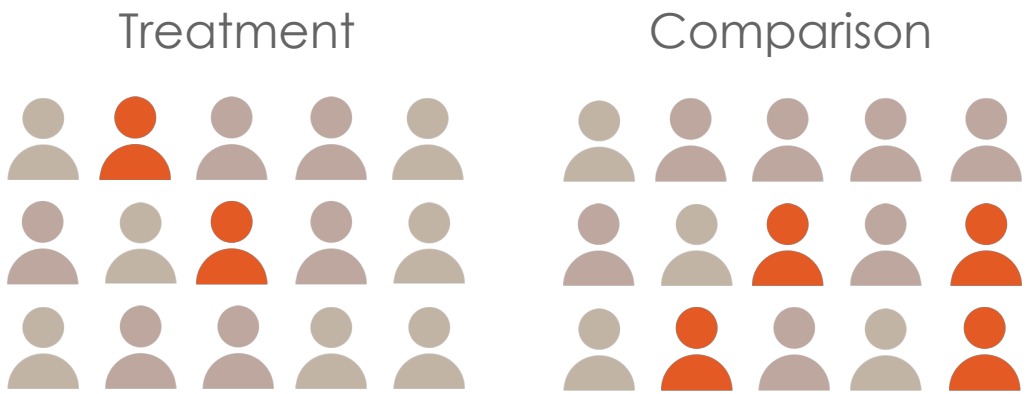
# 3 Key Ideas about Impact

## 1 - Counterfactual




Source: Dan Levy (HKS) and Marc Shotland (JPAL)

## 2 – Comparison group mimics the counterfactual




## 3 - Goal of Impact Evaluations: Attribution

# Impact evaluation methods

1. **Randomized Controlled Trials (RCTs)**

   Also known as:

   - Random Assignment Studies
   - Randomized Field Trials
   - Social Experiments
   - Randomized Trials
   - Randomized Experiments
   - Randomized Controlled Experiments

# Impact evaluation methods

2. **Non- or Quasi-Experimental Methods**
   - Pre-Post
   - Simple Difference
   - Differences-in-Differences
   - Multivariate Regression
   - Statistical Matching
   - Instrumental Variables
   - Regression Discontinuity

# Session Overview

I. Background

II. Why randomize case study

   I. Non-experimental methods

   II. Randomized evaluations

III. Conclusions

# II – Why randomize case study

# Workplace Wellness Programs

**Problem:** Medical spending has risen rapidly over the past several decades, especially in the U.S.

**Proposed solution:** Employer-sponsored workplace wellness programs

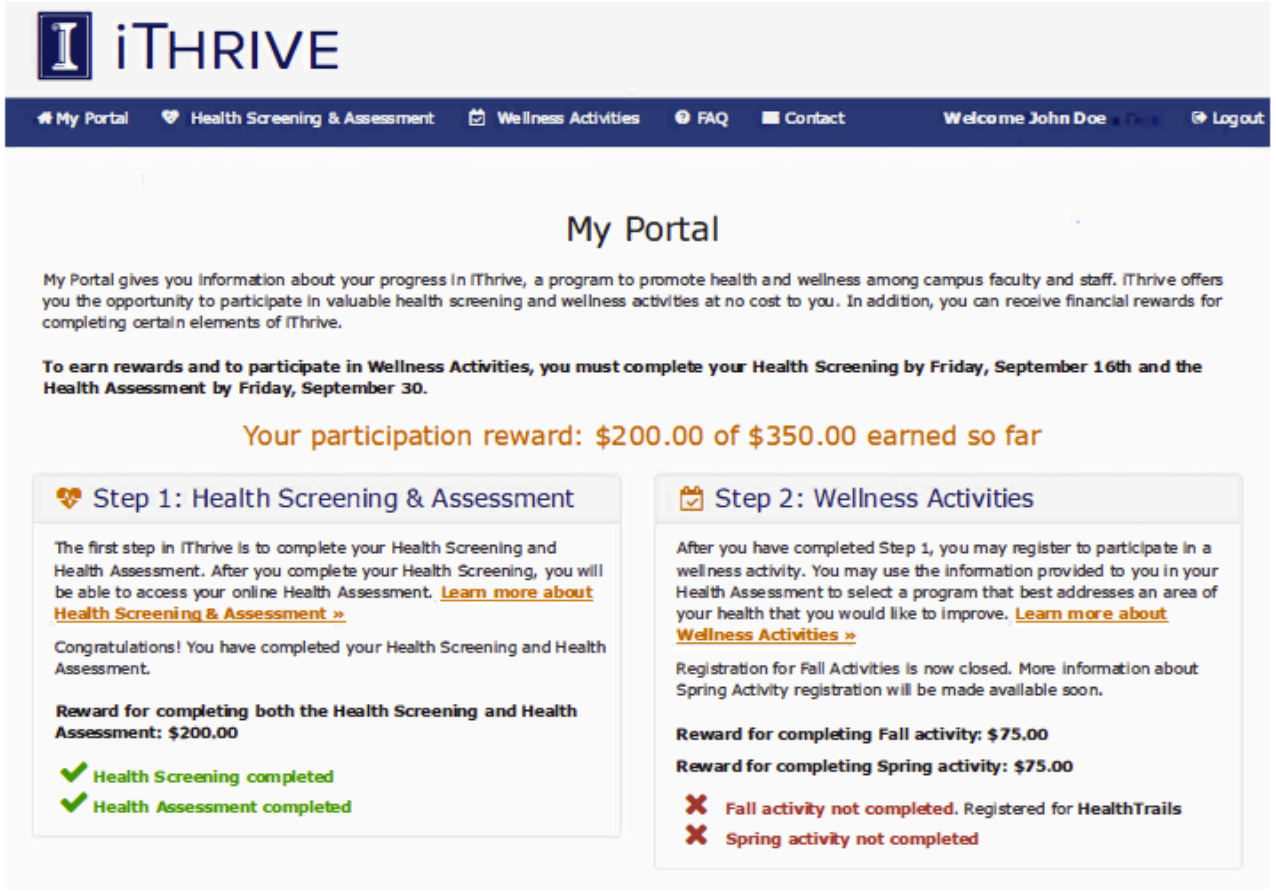
- *Goal: reduce costs by improving employee health*
- Activities include:
  - Health screenings
  - Fitness programs
  - Classes on leading healthy lifestyles

Source: Illinois Workplace Wellness Study

# The Illinois Workplace Wellness Study

In 2016, University of Illinois launched the *iThrive* Wellness Program

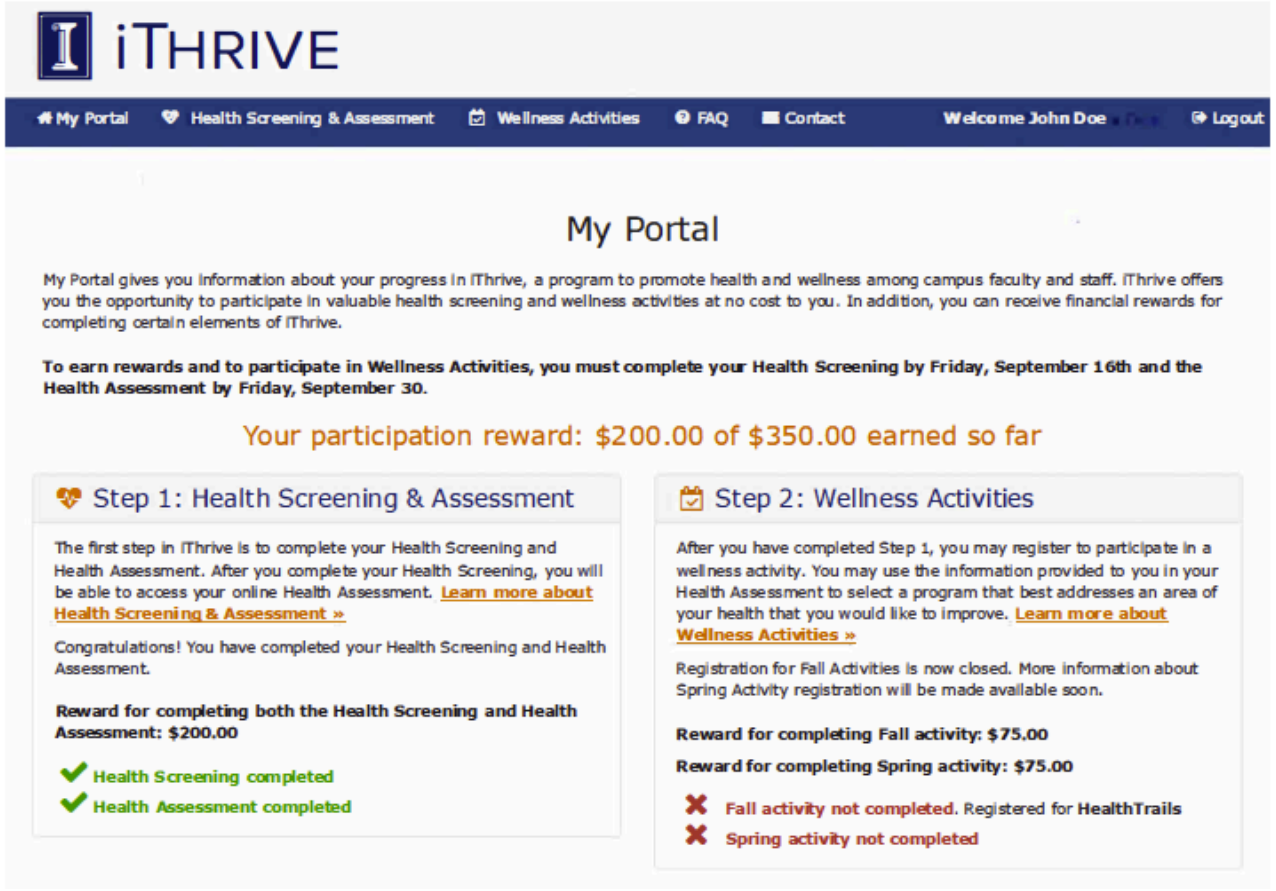
- Biometric screening and health risk assessment
- Wellness activities such as:
  - Exercise classes
  - WeightWatchers
  - Smoking cessation
  - Stress management



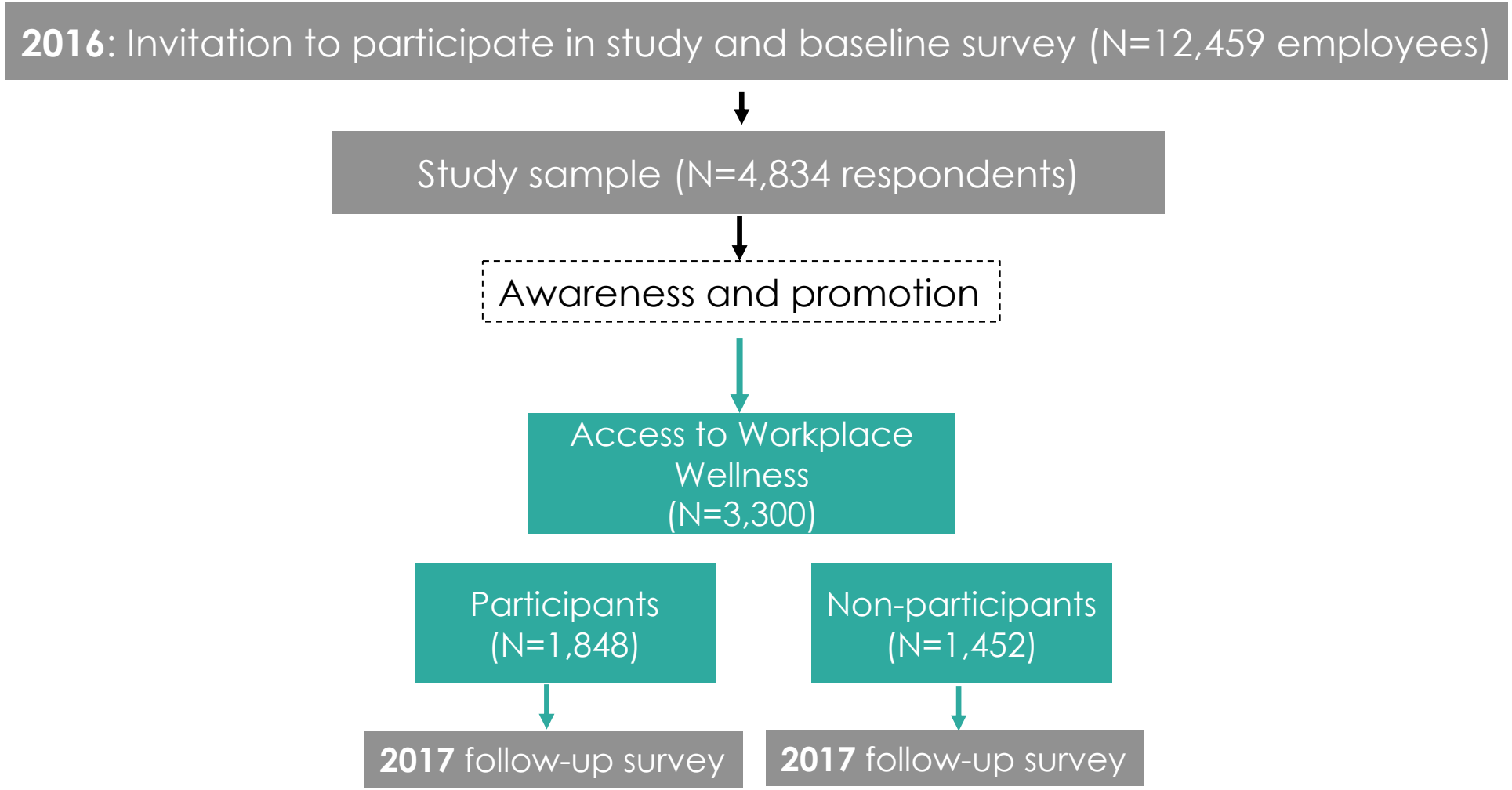
# The Illinois Workplace Wellness Study

Research team focused on two key questions:

- Do wellness programs help employees live healthier lifestyles?
- Are these changes sufficient to lower medical spending?



# The Illinois Workplace Wellness Study

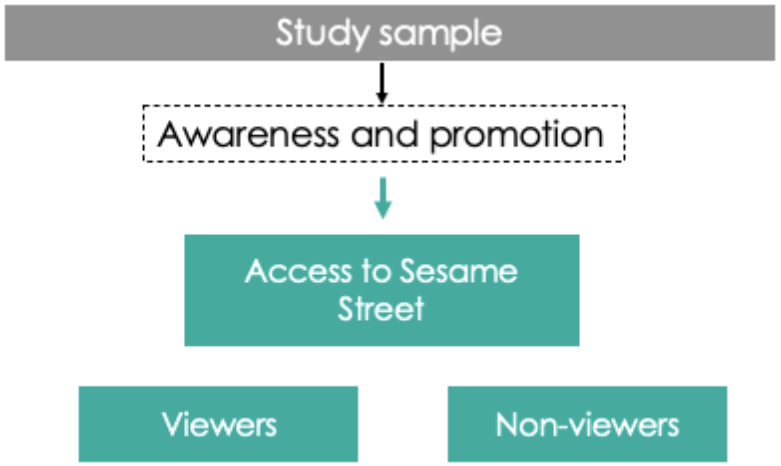


# Outcomes and Data

- Outcomes:
  - Employee fitness:
    - Campus gym visits per month
  - Medical spending:
    - Insurance claims from hospitals, health clinics, and pharmacies
- Background data
  - Age, gender, race, and socio-economic status

# Relevance to other settings

- This is a U.S.-based example, but the takeaways translate to any setting where some people participate in a program, and some do not
- E.g. What is the impact of Sesame Street on child literacy and numeracy?




- Challenge is the same: find a valid counterfactual for participants / viewers

# Session Overview

I. Background

II. Why randomize case study

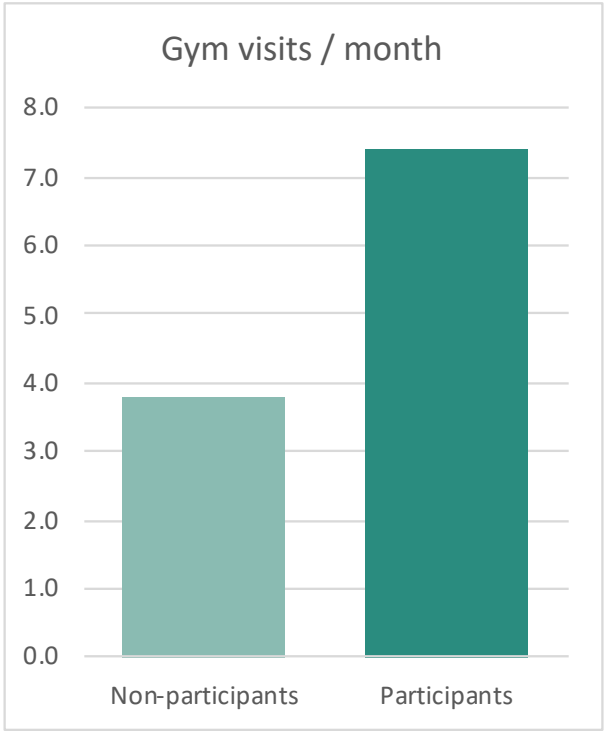
   I. Non-experimental methods

   II. Randomized evaluations

III. Conclusions

# Simple difference

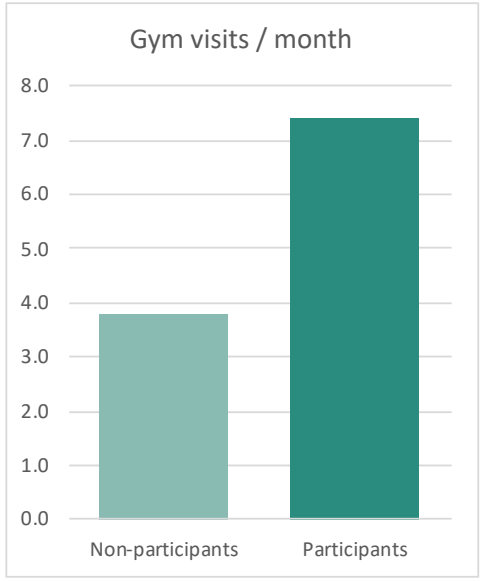
Compare employees who participated in the program to those that did not

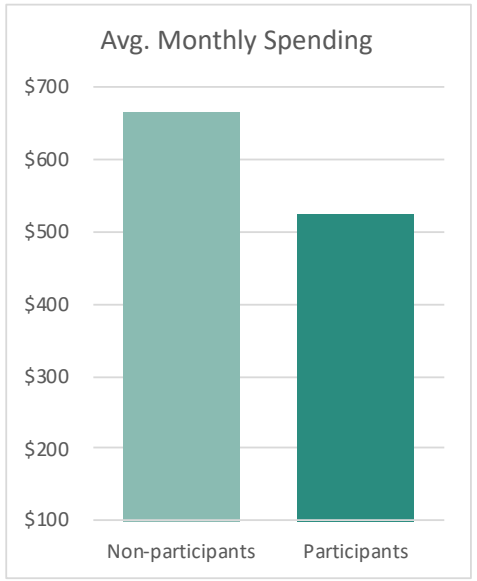



# Simple Difference

Counterfactual: Non-participants' frequency of gym visits and levels of medical spending

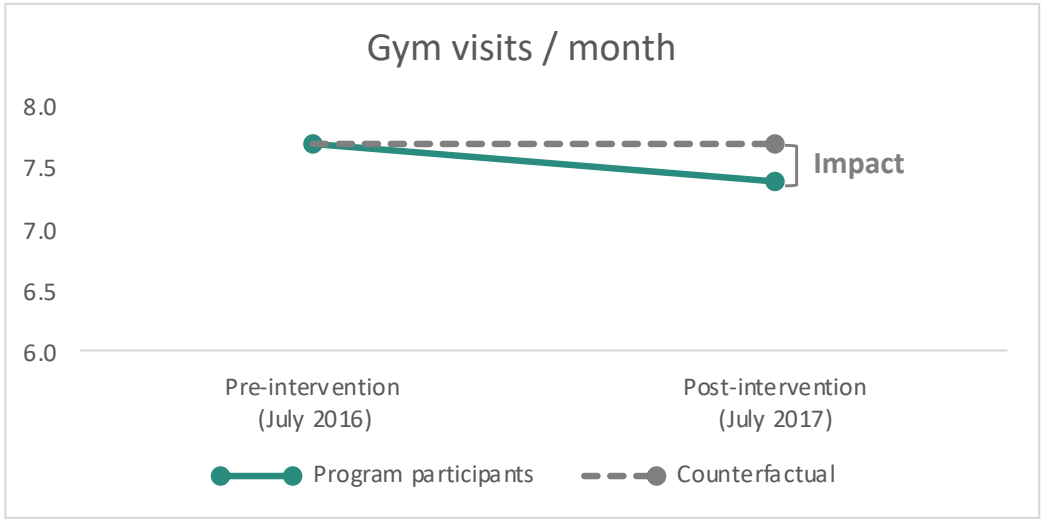
Key assumption: Participants' would have had the same levels of these outcomes as non-participants *if they had not received the program*

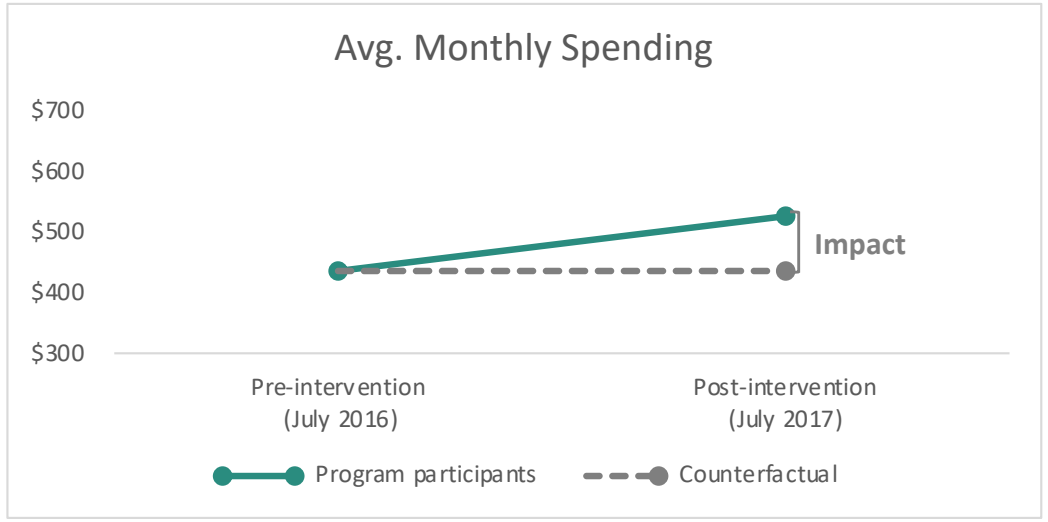


# Pre-Post (Before vs. After)

Compare participants' outcomes before the program to their outcomes after the program

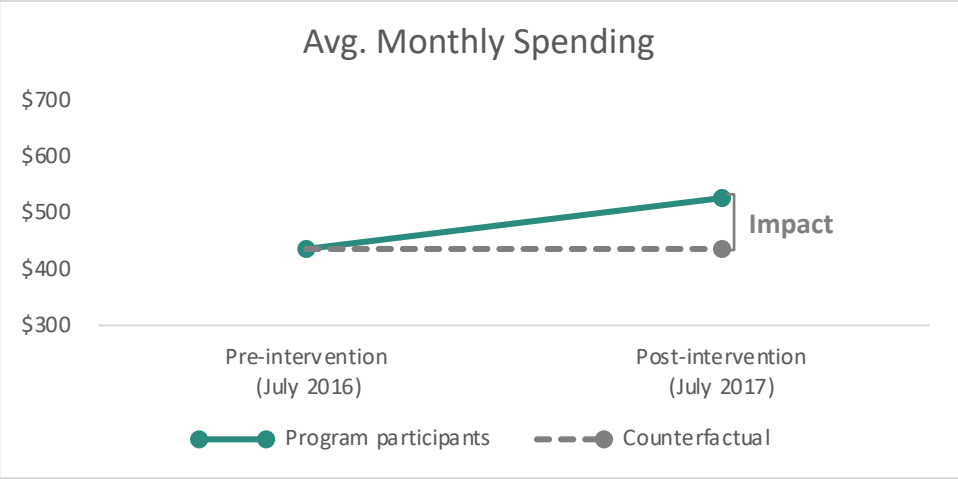


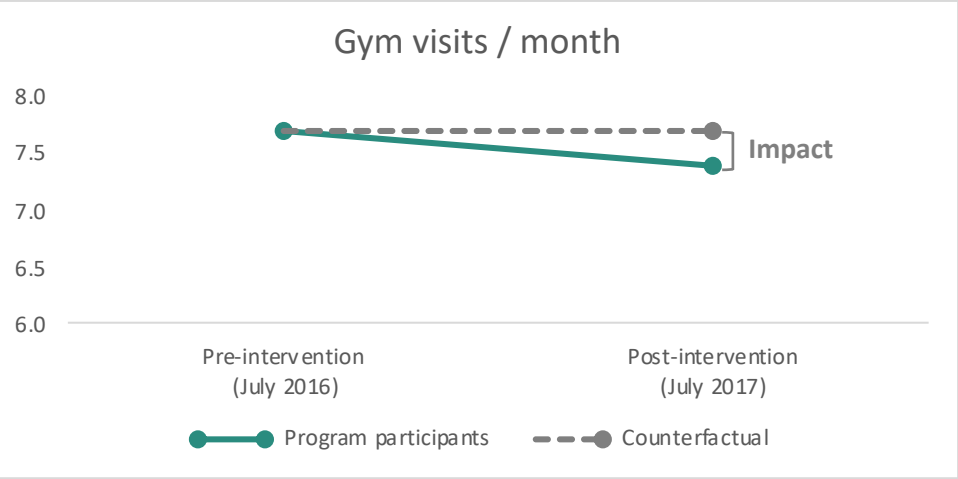



# Pre-Post (Before vs After)

Counterfactual: Participants' average levels of fitness and spending, and *before the program*

Assumptions: Participants' fitness, spending, and productivity would not have changed over time in the absence of the program



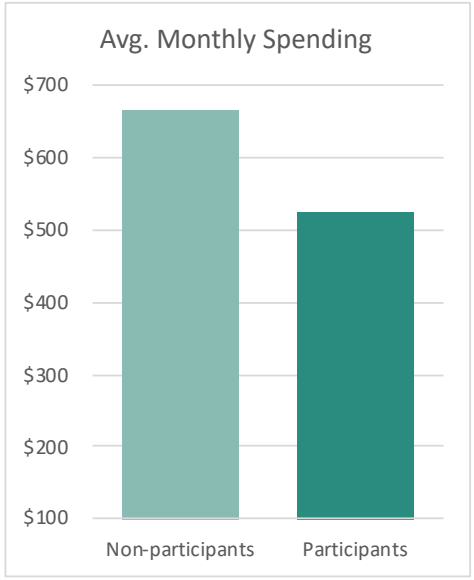
# Breakout Discussion I

# Breakout Discussion I – 12 minutes

- Housekeeping:
  - Please turn your video on
  - When speaking, turn your audio on as well
- Agenda
  - Simple difference comprehension questions
  - Pre / post comprehension questions
  - Any questions so far

# Which of these scenarios would make the ***Simple Difference* method** misleading? (select all that apply)

- A: Healthcare costs tend to increase overtime
- B: Participants *live closer to campus* than non-participants
- C: Participants are *younger* than non-participants



# Simple Difference - Recap

Counterfactual: Non-participants' frequency of gym visits and levels of spending

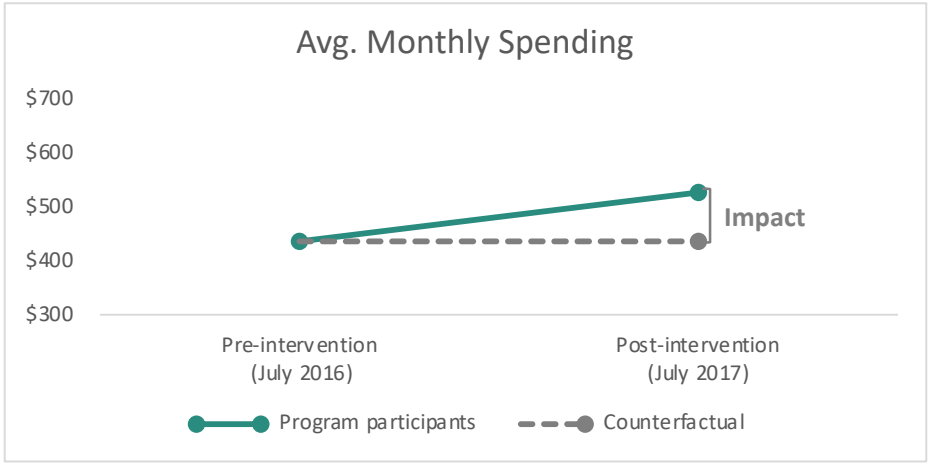
Key assumption: Participants' would have the same levels of these outcomes as non-participants *if they had not received the program*


What might threaten this assumption: Any systematic difference between participants and non-participants that influences our outcomes

Statisticians call this ***selection bias***, because those who "select in" to a program may be different from those who do not in terms of their pre-program outcomes

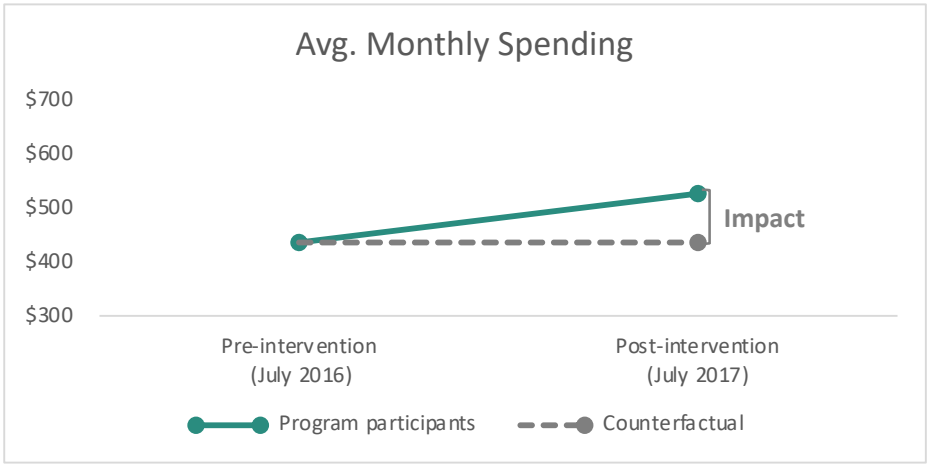
# Which of these scenarios would make the ***Pre-Post*** method misleading? (select all that apply)

- A: Healthcare costs tend to increase overtime
- B: Participants *live closer to campus* than non-participants
- C: Participants are *younger* than non-participants



# Pre-Post (Before vs After) - Recap

Counterfactual: Participants' average levels of fitness, spending, and productivity *before the program*



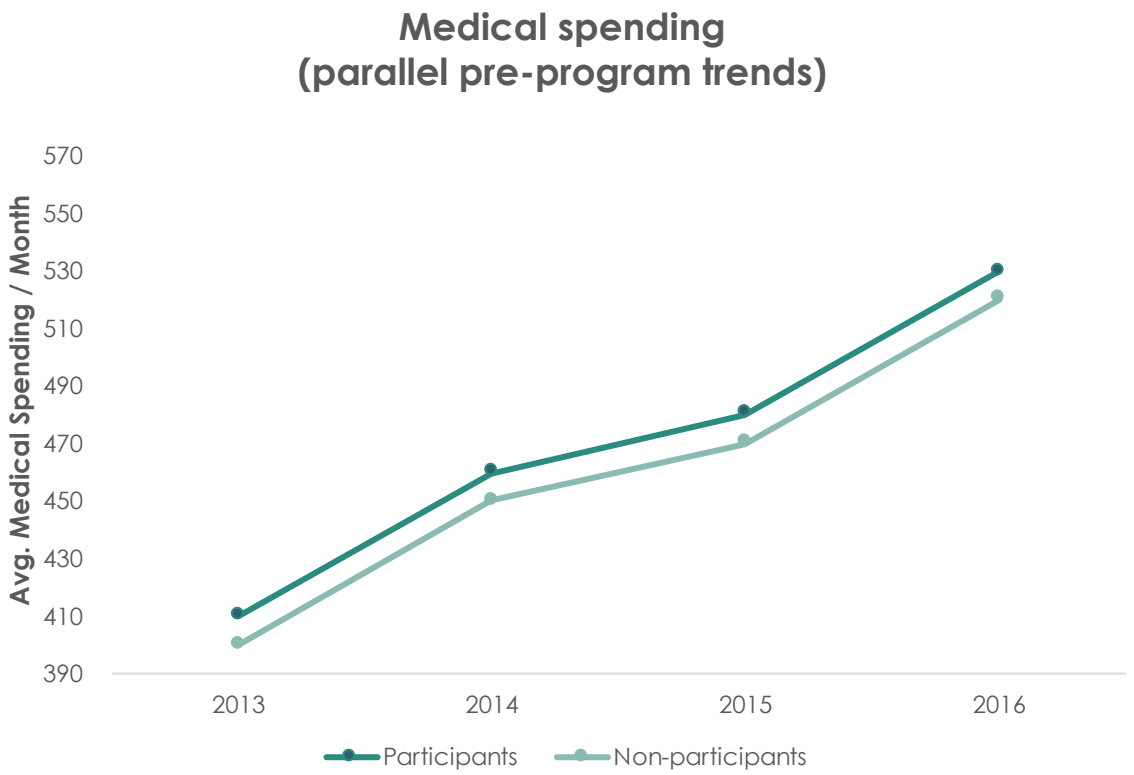

Assumptions: Participants' fitness, spending, and productivity would not have changed over time in the absence of the program

What might threaten this assumption: Any factor that influences these outcomes overtime

# Any questions on material covered thus far?

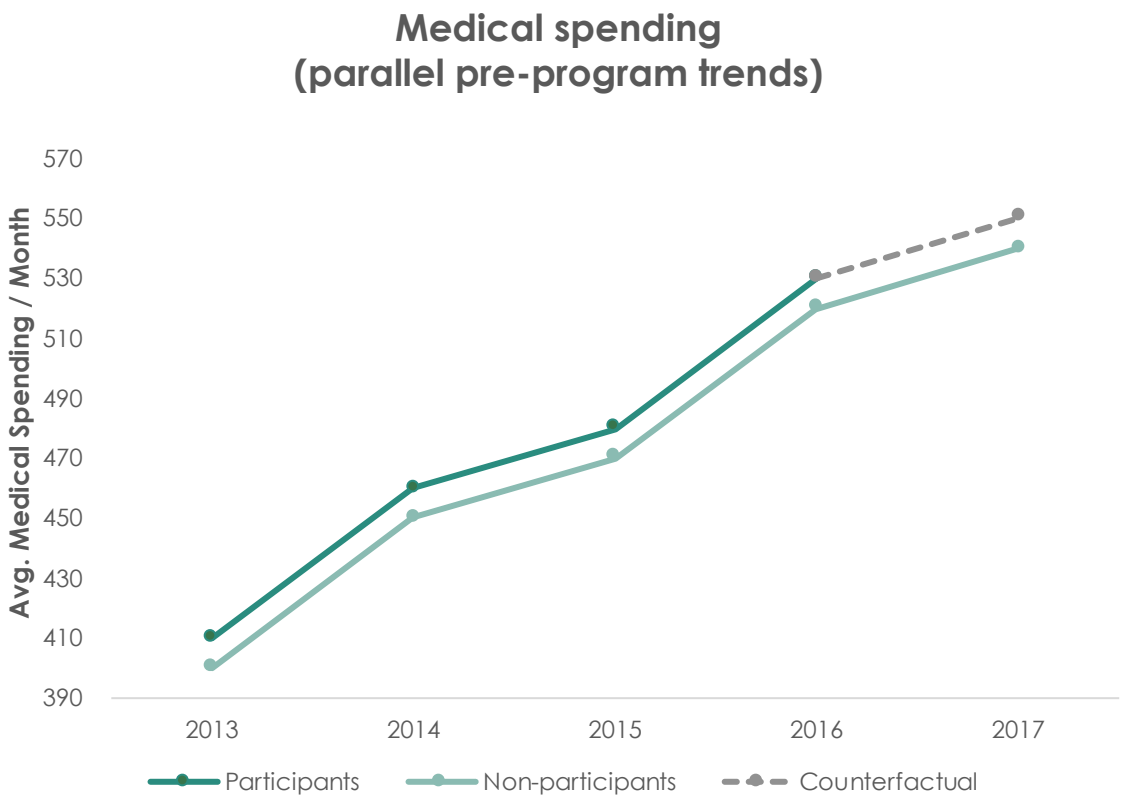
# End of Breakout I

# Difference-in-differences



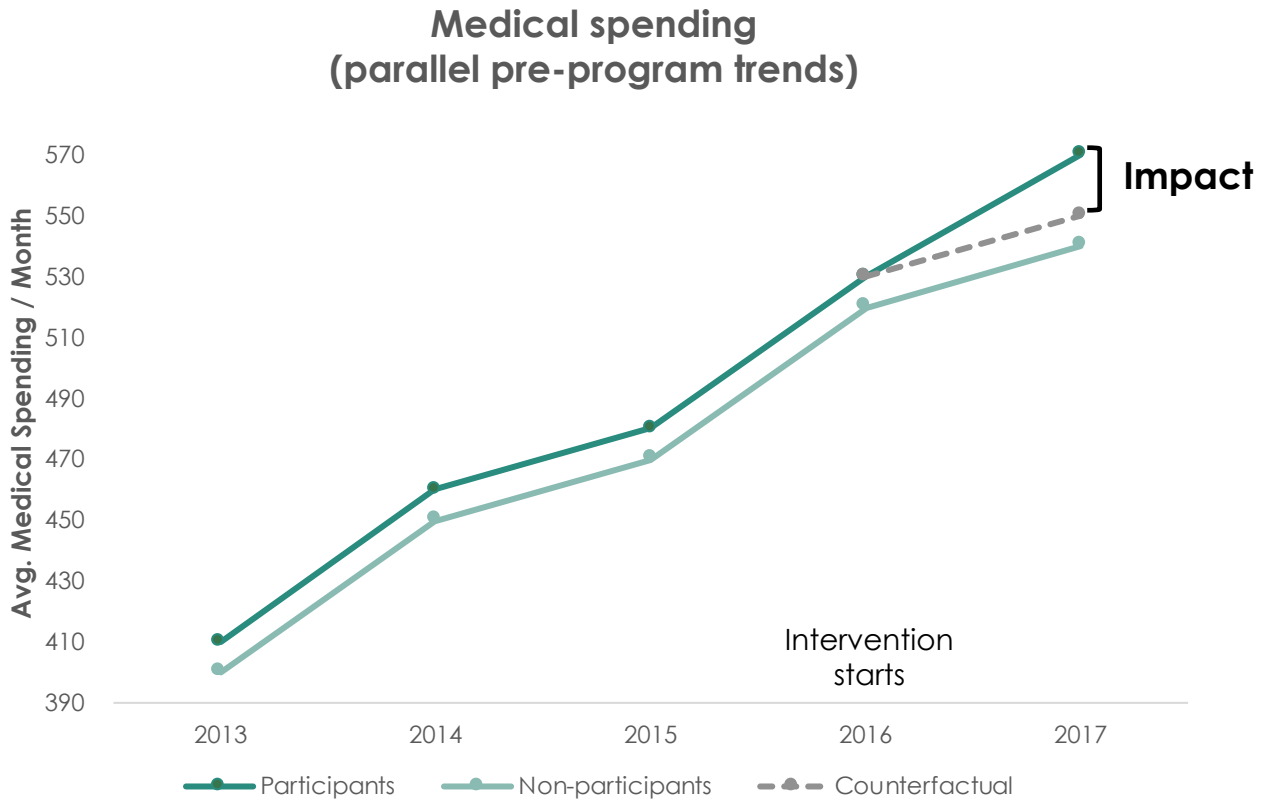

*Note:* Figure is illustrative and not based on actual study data

# Difference-in-differences



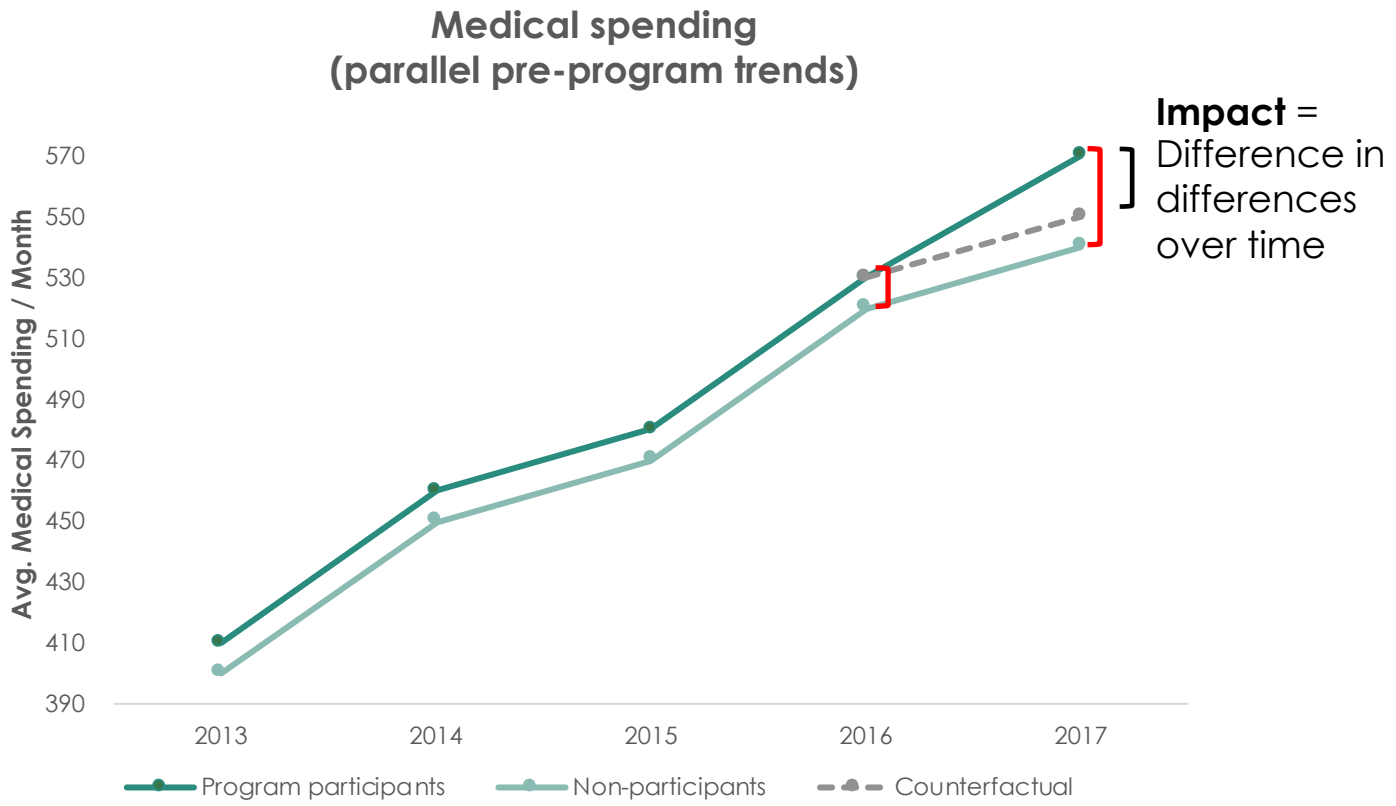

*Note:* Figure is illustrative and not based on actual study data

# Difference-in-differences




*Note:* Figure is illustrative and not based on actual study data

# Difference-in-differences

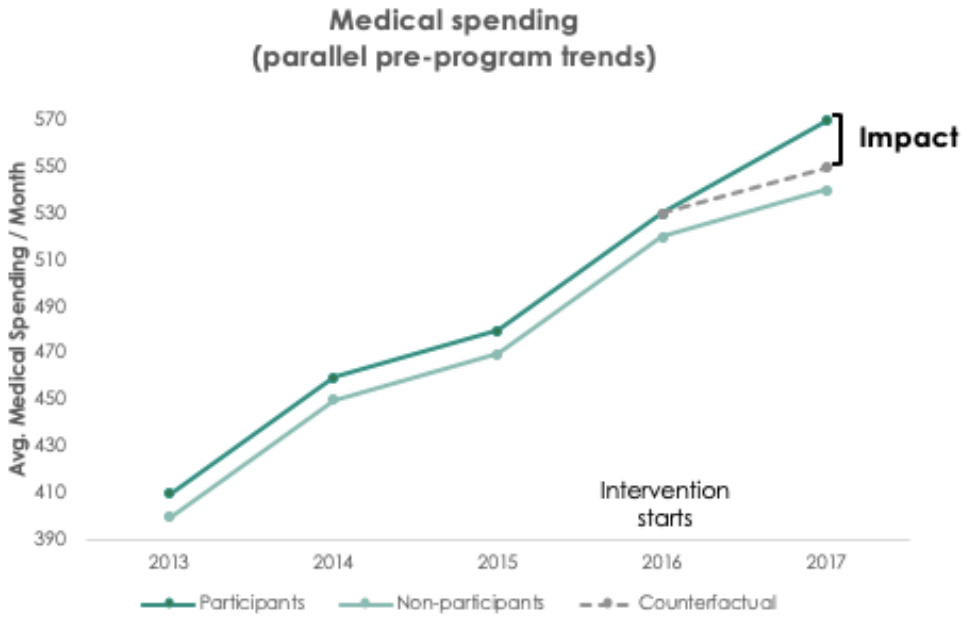



*Note:* Figure is illustrative and not based on actual study data

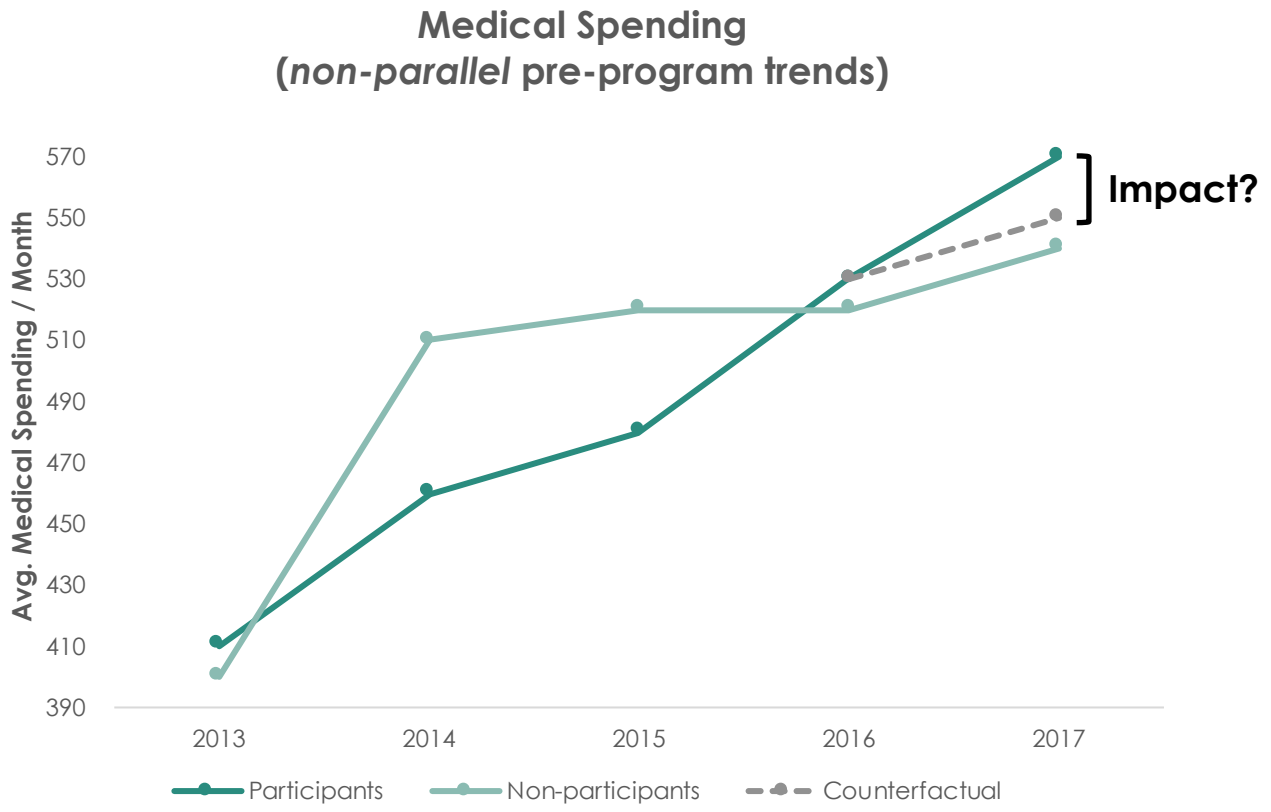
# Difference-in-differences

What's the counterfactual? Pre / post change over time in medical spending among *non-participants*

Assumptions: Absent the program, participants and non-participants would have the same trajectory over time with respect to medical spending ("**parallel trends**" assumption)



# Difference-in-differences

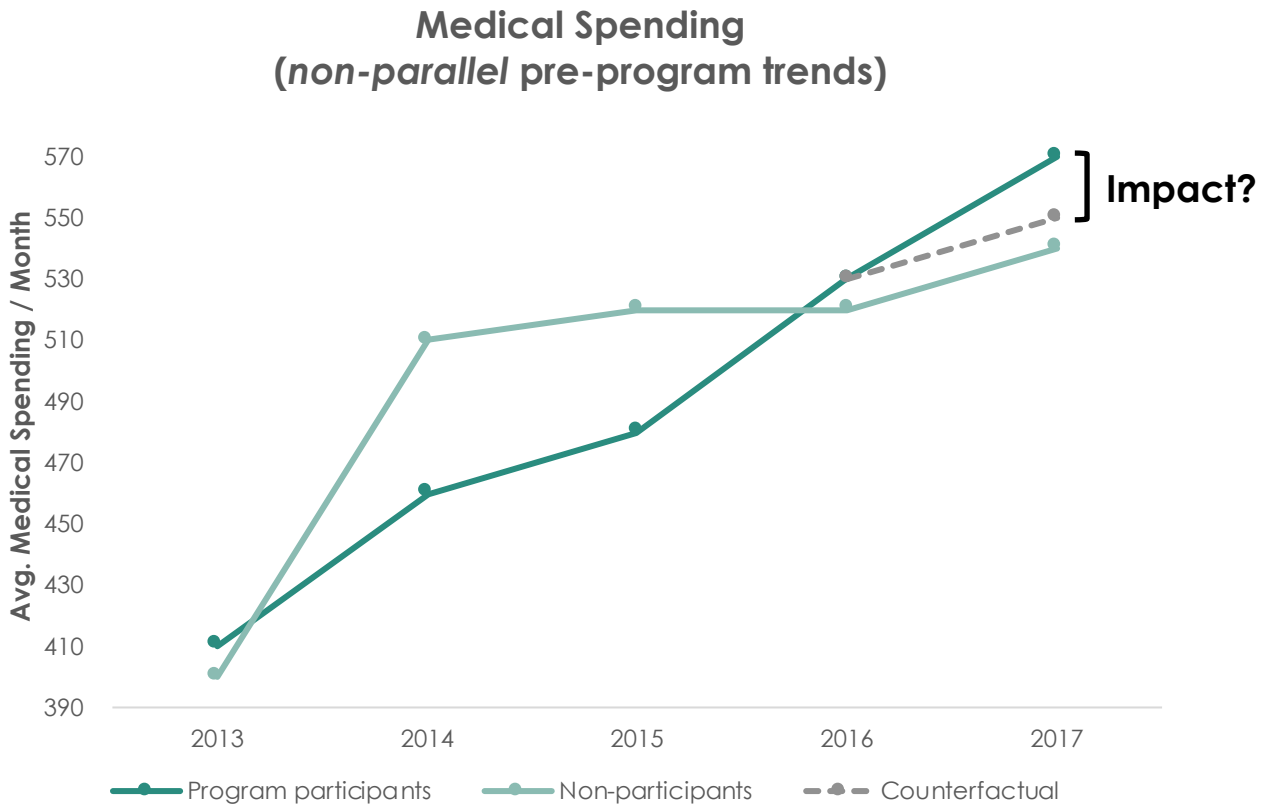



*Note:* Figure is for instructional purposes only and not based on actual study data

# Which of these scenarios would lead to non-parallel trends? (select all that apply)

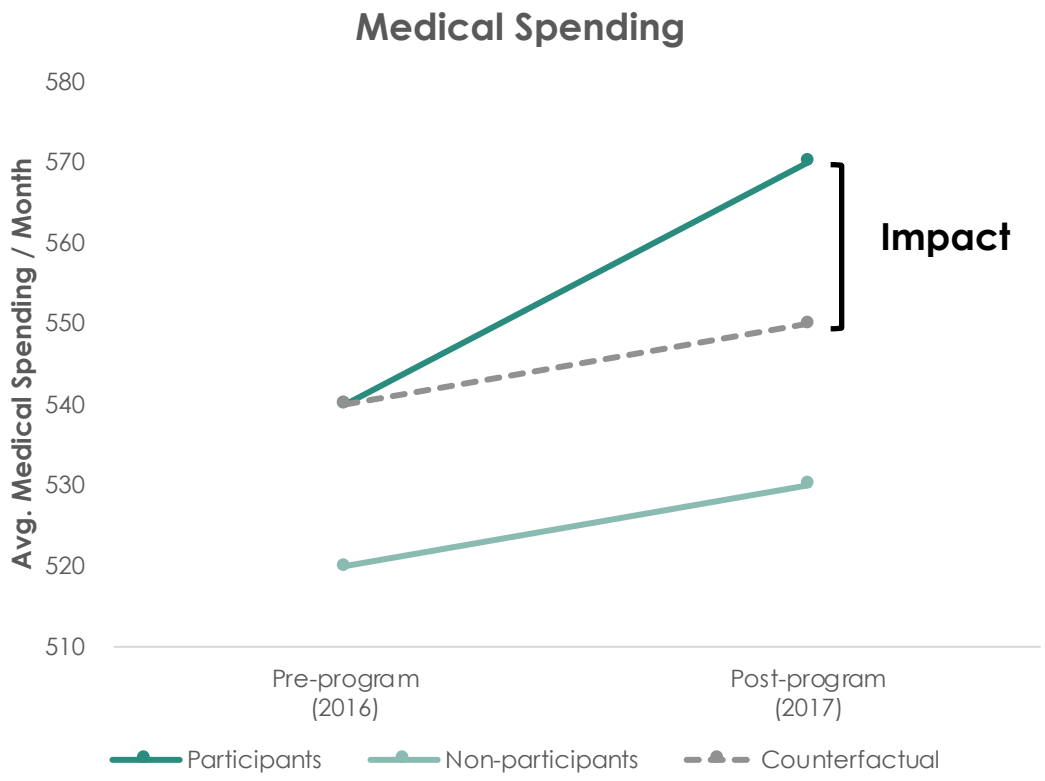
- A: Healthcare costs tend to increase overtime
- B: Participants are *older* than non-participants
- C: An effective but expensive drug for a common chronic condition among older adults enters the market in 2014
- D: Options B and C at once

# Difference-in-differences




*Note:* Figure is for instructional purposes only and is not based on actual study data

# Difference-in-differences




*Note:* Figure is illustrative and not based on actual study data

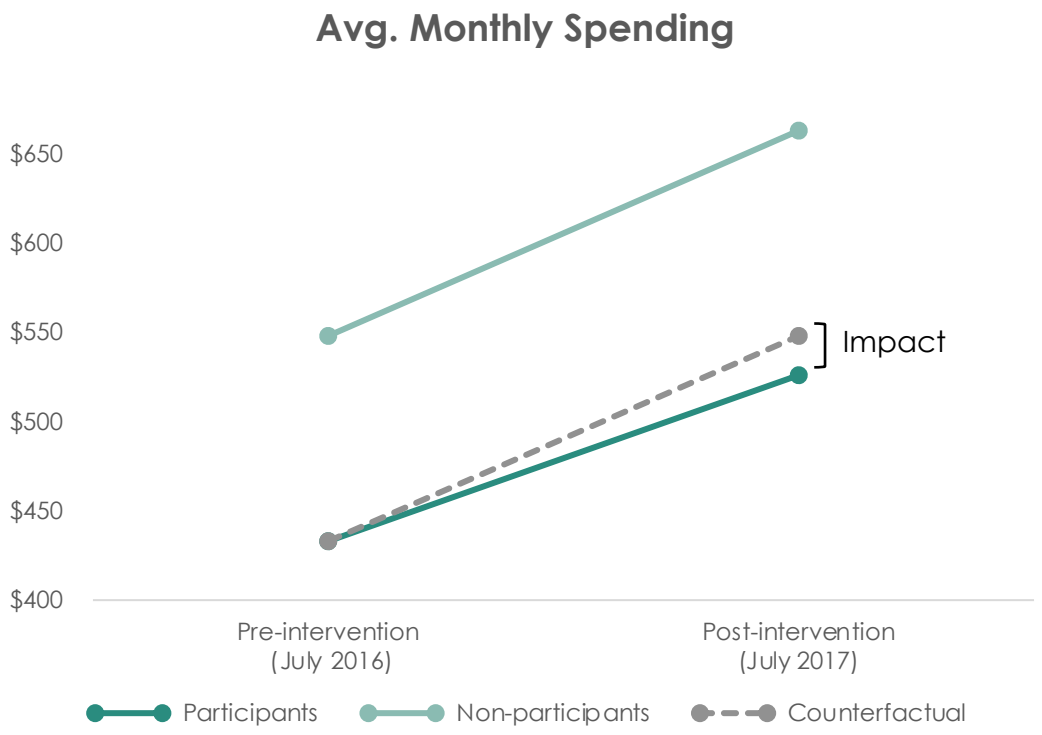
# Difference-in-differences

What's the counterfactual? Pre / post change over time in medical spending among *non-participants*

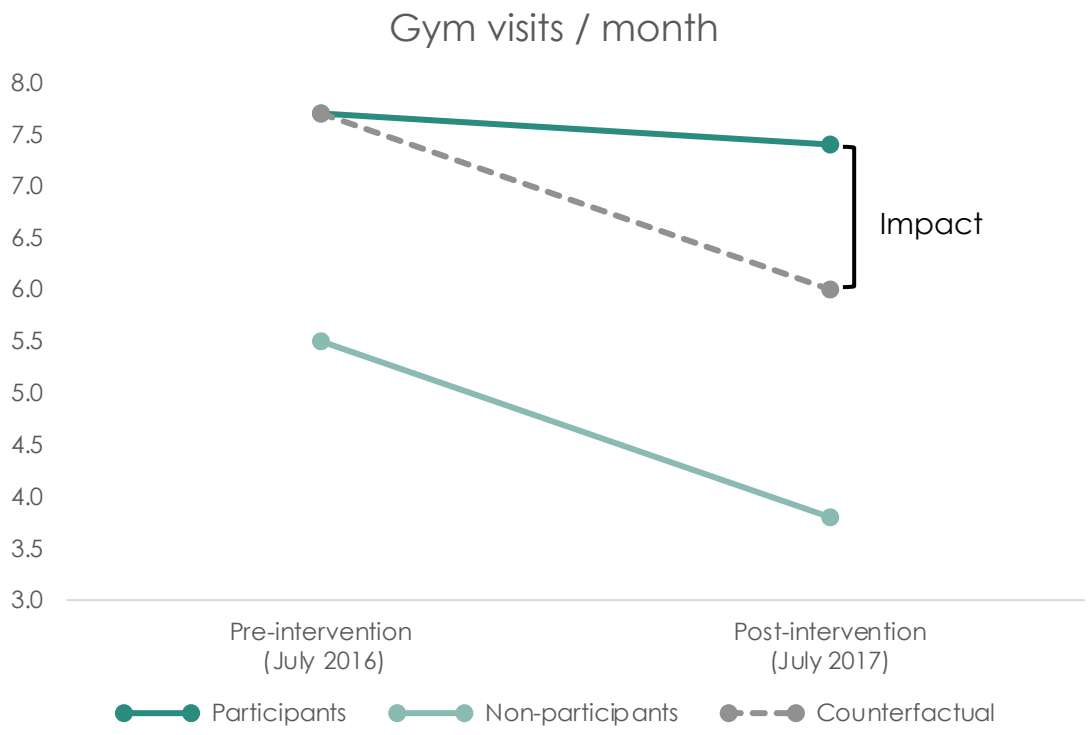
Assumptions: Absent the program, participants and non-participants would have the same trajectory over time with respect to medical spending ("**parallel trends**" assumption)

What might threaten this assumption: Any change over time that disproportionately impacts either group

# Difference-in-differences - results



# Difference-in-differences - results



# Recap of results so far

| Method | Gym visits / month | Medical Spending |
|---|---|---|
| (1) Pre-Post | 3.5** | -$137** |
| (2) Simple Difference | -0.4 | $100** |
| (3) Difference-in-Differences | 1.34** | -$9.6 |

# Statistical Control - Matching

**Motivation**: Before the intervention, participants were different from non-participants in various ways

**Table 1 - Pre-intervention characteristics by participation status, before matching**

| | Non-participants | Participants | Difference | N |
|---|---|---|---|---|
| Avg. monthly spending (pre-intervention) | $527 | $423 | $103** | 3300 |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

Notes: ***, **, and * indicate significance at the p-value < .01, .05, and .10 levels. Sample sizes vary across outcomes due to missing data.

# Statistical Control - Matching

**Motivation**: Before the intervention, participants were different from non-participants in various ways

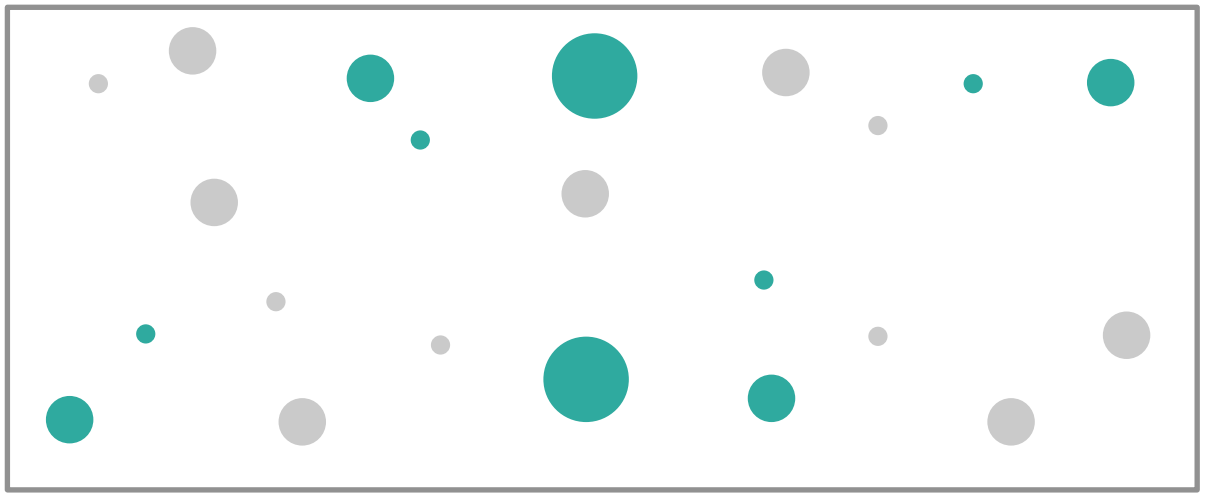
**Table 1 - Pre-intervention characteristics by participation status, before matching**

| | Non-participants | Participants | Difference | N |
|---|---|---|---|---|
| Avg. monthly spending (pre-intervention) | $527 | $423 | $103** | 2188 |
| Gym visits per month (pre-intervention) | 5.6 | 7.7 | -2.2** | 3300 |
| Male | 46% | 40% | 6%** | 3300 |
| High salary | 48% | 51% | 3%* | 3300 |
| Faculty | 23% | 18% | 5%* | 3300 |

Notes: ***, **, and * indicate significance at the p-value < .01, .05, and .10 levels. Sample sizes vary across outcomes due to missing data.

# Statistical Control - Matching

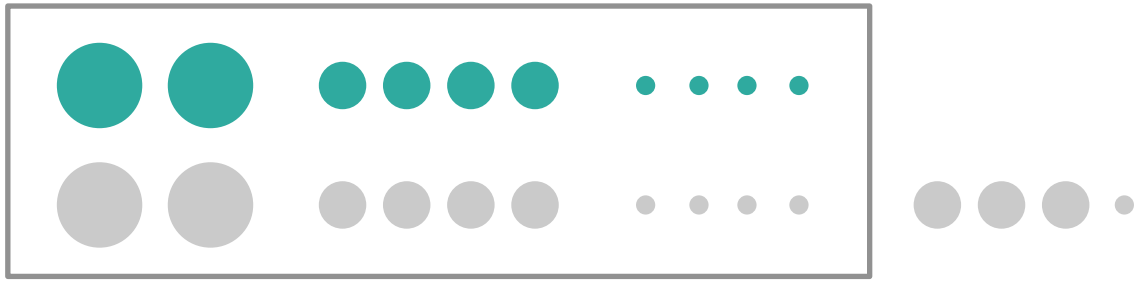
Study sample with varying characteristics



Treatment

Comparison

Study group with matching



# Statistical Control - Matching

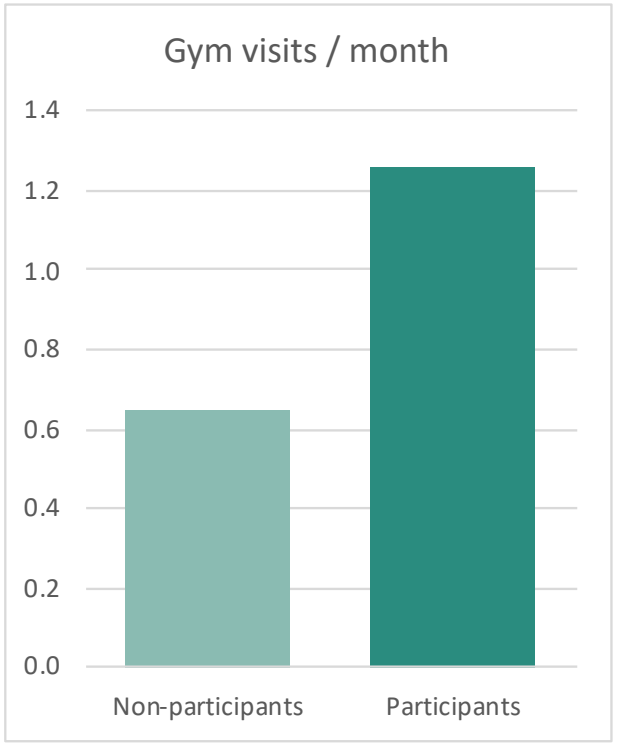
After matching, participants look more similar to non-participants, but sample size is smaller

**Table 2 - Pre-intervention characteristics by participation status, after matching**

| | Non-participants | Participants | Difference | N |
|---|---|---|---|---|
| Avg. monthly spending (pre-intervention) | $203 | $184 | $19 | 1109 |
| Gym visits per month (pre-intervention) | 0.47 | 0.33 | 0.14 | 1109 |
| Male | 0.45 | 0.39 | 6%** | 1109 |
| High salary | 48% | 48% | 0% | 1109 |
| Faculty | 11% | 8% | 3% | 1109 |

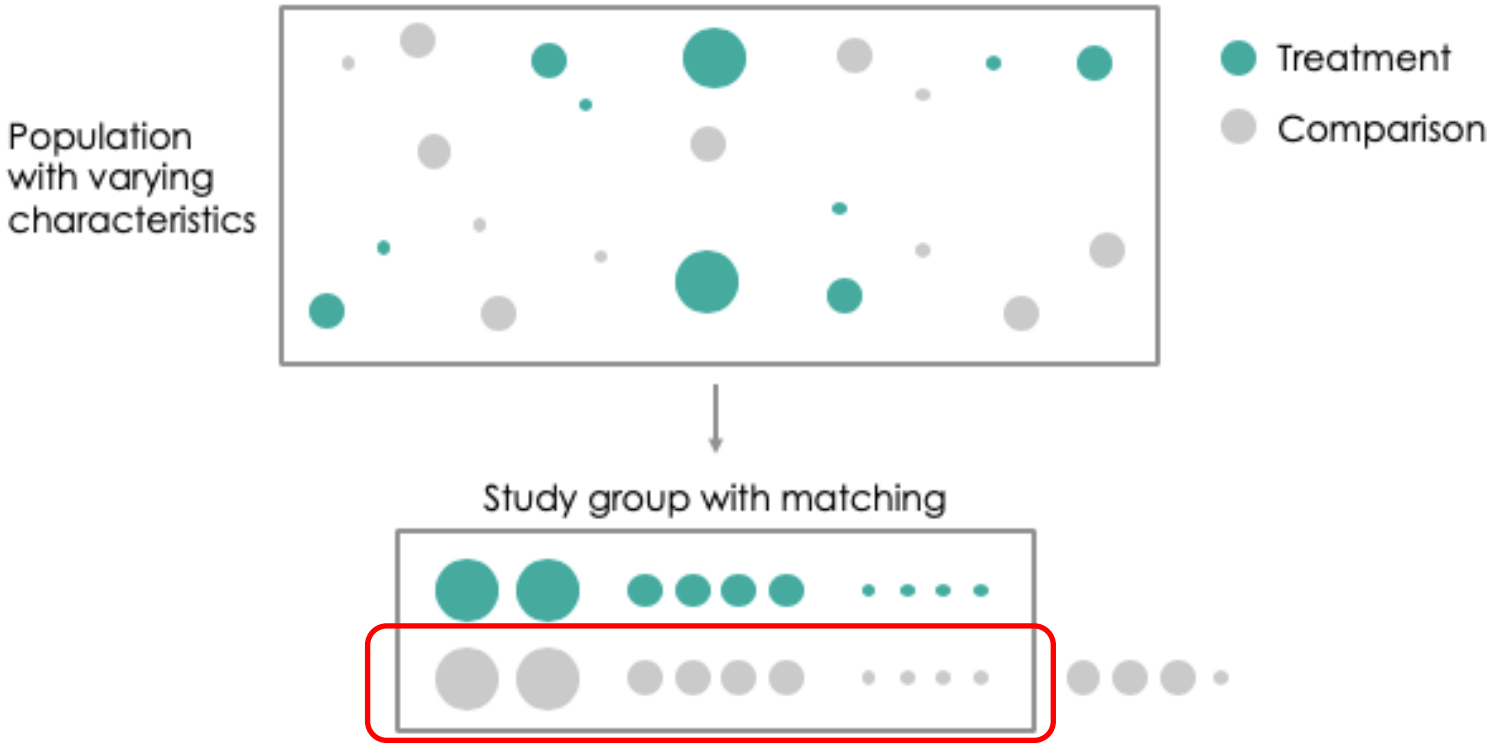
Notes: ***, **, and * indicate significance at the p-value < .01, .05, and .10 levels. Sample sizes vary across outcomes due to missing data.

# Statistical Control – Matching




# Statistical Control - Matching

Counterfactual: Gym visits, levels of spending, and productivity among non-participants for whom there was a comparable participant match



# Statistical Control - Matching

Counterfactual: Gym visits, levels of spending, and productivity among non-participants for whom there was a comparable participant match

Key assumption: Participants' would have had the same levels of these outcomes as their non-participant matches *if they had not received the program*

# Breakout Discussion II

# Breakout Discussion II – 8 minutes

- Housekeeping:
  - Please turn your video on
  - When speaking, turn your audio on as well
- Agenda
  - Statistical Matching comprehension questions
  - Any outstanding questions

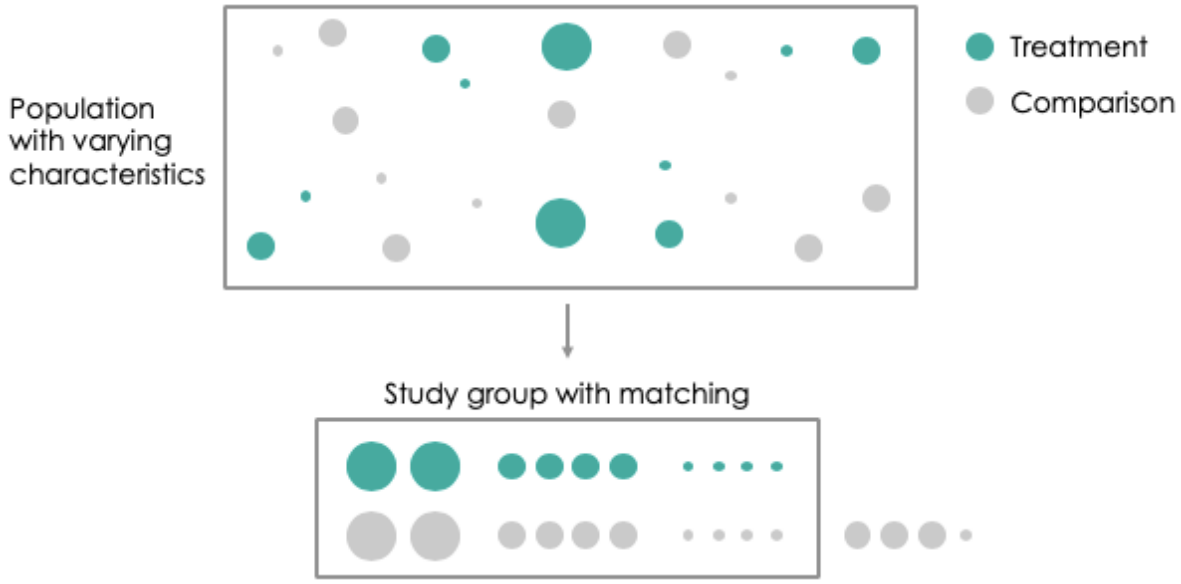
# Which of these scenarios would make participant vs. non-participant comparisons within the matched sample misleading?

- A: Healthcare costs tend to increase overtime
- B: Participants are *older* than non-participants
- C: Participants are more *intrinsically motivated* to improve their health than non-participants

# Statistical Matching - Recap

Counterfactual: Gym visits and medical spending among non-participants matches

Key assumption: Participants' would have had the same levels of these outcomes as their non-participants matches *if they had not received the program*




What might threaten this assumption: Any systematic difference between participants and non-participants that we can't measure (or forget to measure), that also influences outcomes

# Any questions on material covered thus far?

# End of Breakout II

# Session Overview

I. Background

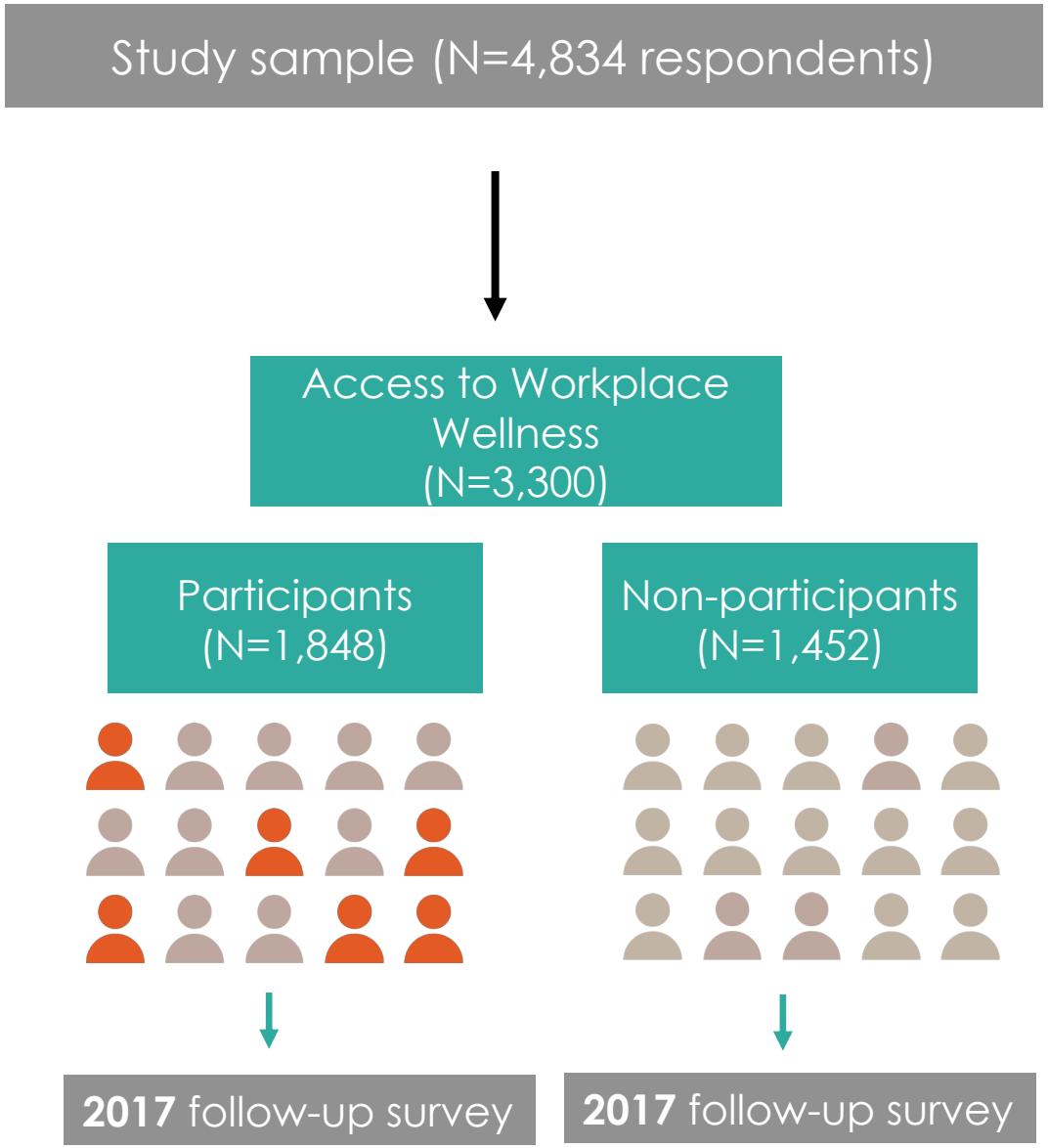
II. Why randomize case study

    I. Non-experimental methods

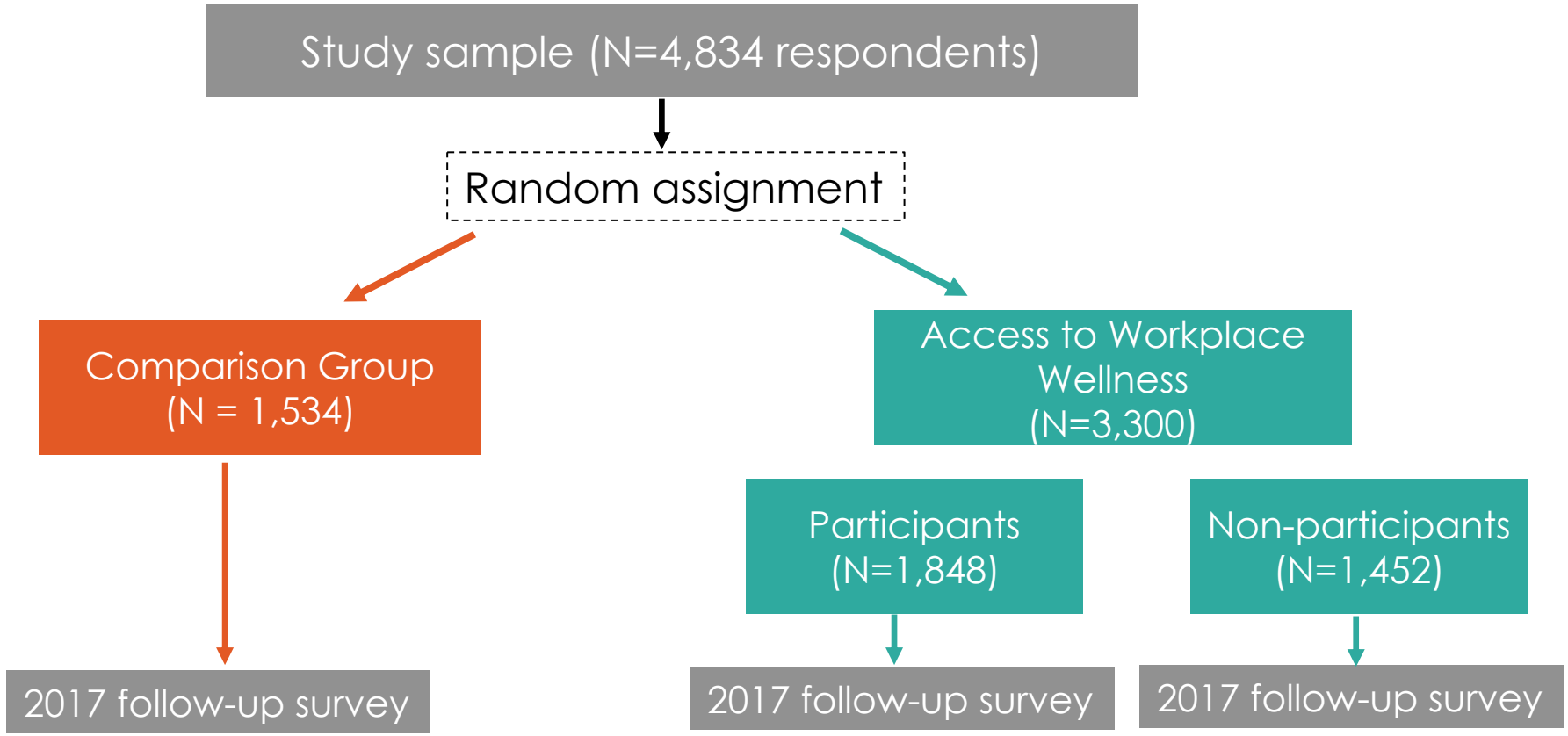
    II. Randomized evaluations

III. Conclusions

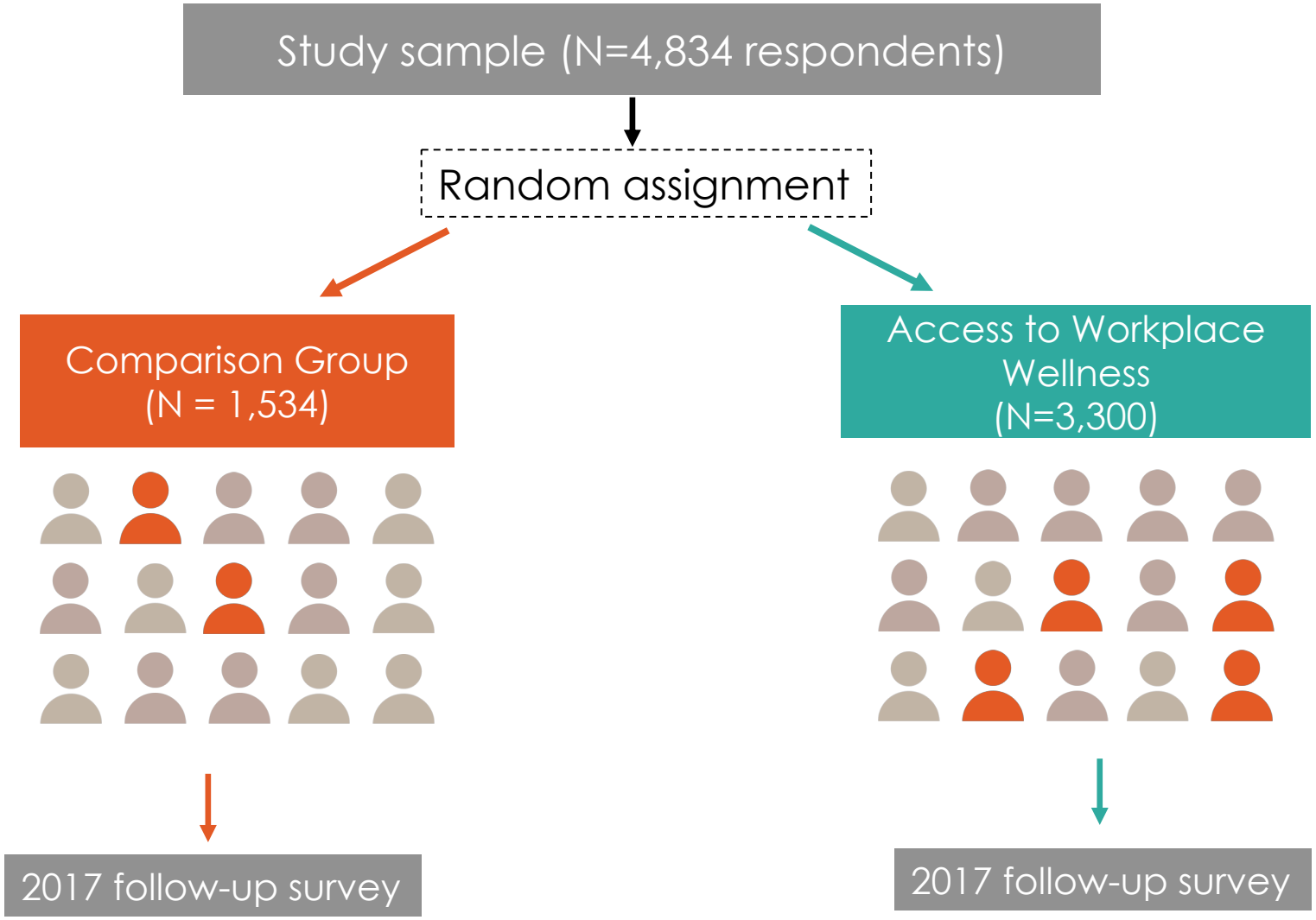
# Problem so far:



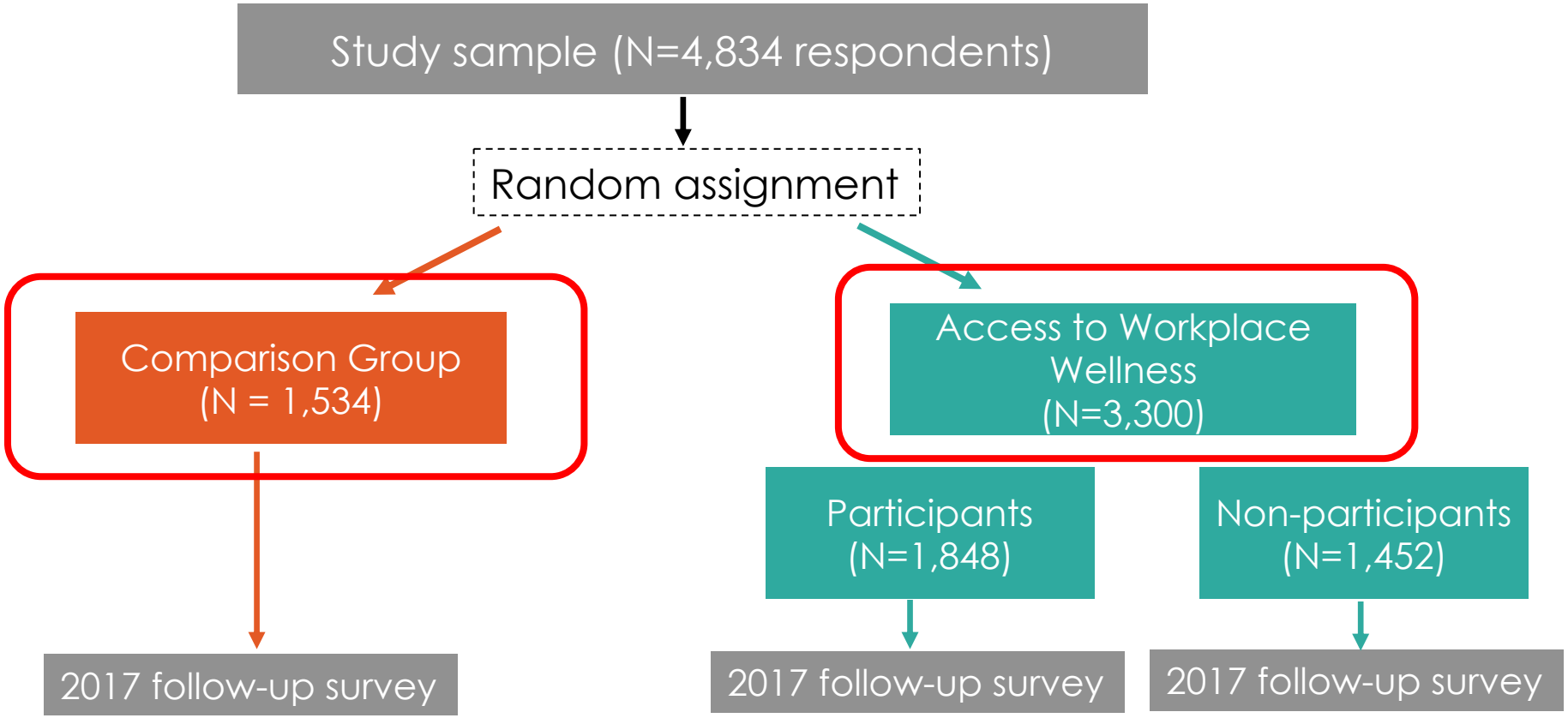
# Randomized Evaluation



# Randomized Evaluation



# Randomized Evaluation



# Randomized Evaluation

**Key advantage:** Because members of the groups (treatment and comparison) do not differ systematically at the outset of the evaluation,

any difference that subsequently arises between them can be attributed to the program rather than to other factors.

Comparison Group

Access to Wellness Program

# Workplace Wellness – treatment vs. comparison at baseline (2016)

**Table 2 - Pre-intervention characteristics by participation status, after matching**

| | Offered program | Not offered program | Difference | N |
|---|---|---|---|---|
| Avg. monthly spending (pre-intervention) | $478 | $496 | $18 | 4,834 |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

Notes: ***, **, and * indicate significance at the p-value < .01, .05, and .10 levels. Sample sizes vary across outcomes due to missing data.

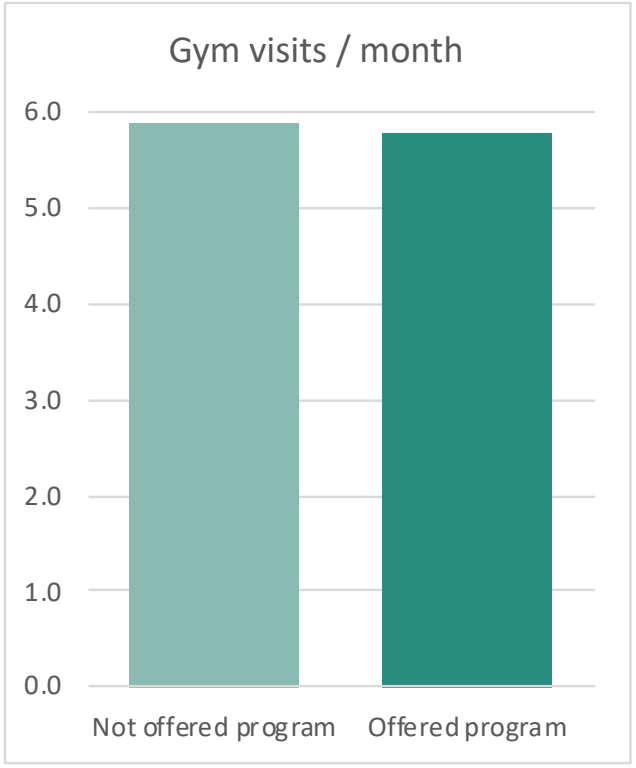
# Workplace Wellness – treatment vs. comparison at baseline (2016)

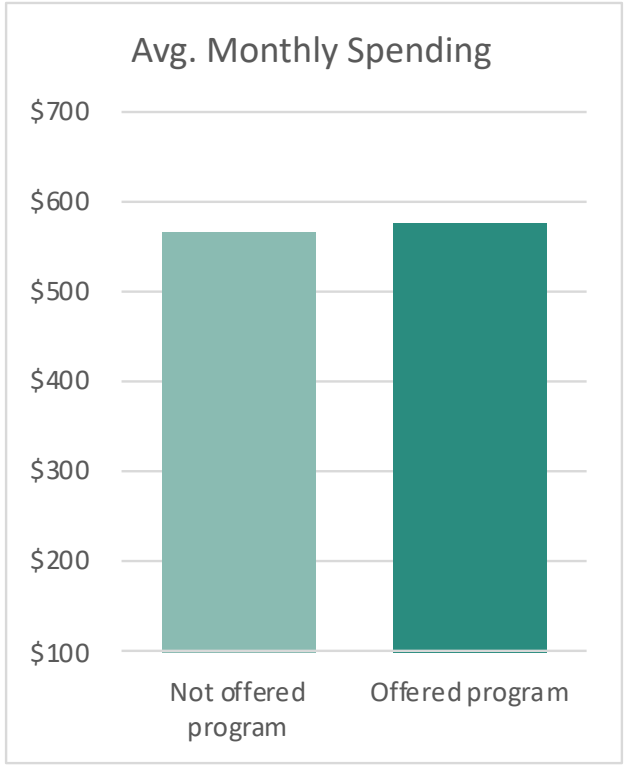
**Table 2 - Pre-intervention characteristics by participation status, after matching**

| | Offered program | Not offered program | Difference | N |
|---|---|---|---|---|
| Avg. monthly spending (pre-intervention) | $478 | $496 | $18 | 4,834 |
| Gym visits per month (pre-intervention) | 6.8 | 7.2 | 0.4 | 4,834 |
| Male | 43% | 43% | 0% | 4,834 |
| High salary | 50% | 50% | 0% | 4,834 |
| Faculty | 19% | 19% | 0% | 4,834 |

Notes: ***, **, and * indicate significance at the p-value < .01, .05, and .10 levels. Sample sizes vary across outcomes due to missing data.

# Randomized Evaluation - Results



# Comparison of results across methods

| Method | Gym visits / month | Medical Spending |
|---|---|---|
| (1) Pre-post | 3.5** | -$137** |
| (2) Simple difference | -0.4 | $100** |
| (3) Difference-in-differences | 1.34** | -$9.6 |
| (4) Matching | 0.61 | -$146 |
| (5) Randomized evaluation | 0.06 | $10 |

*Note:* *, **, and *** indicate statistical significance at the p-value < .10, .05 and .01.

# Breakout Discussion III

# Breakout Discussion III – 10 minutes

Any outstanding questions on the randomized evaluation method?

Think of a program from your area of work

- Why would it be necessary to run a randomized evaluation of this program? How might participants differ from non-participants?
- How would you design a randomized evaluation to evaluate this program?

# End of Breakout III

# Session Overview

I. Background

II. Why randomize case study

    I. Non-experimental methods

    II. Randomized evaluations

III. Conclusions

# IV – CONCLUSIONS

# Conclusions – Why Randomize?

- There are many ways to estimate a program's impact
- This lecture highlights the advantages one: randomized evaluations
  - **Conceptual argument:** If properly designed and conducted, randomized evaluations are the most credible method to estimate the impact of a program
  - **Empirical argument:** Different methods can generate different impact estimates
- When randomized evaluations are impractical, non-experimental methods may be the best option. But being clear about the counterfactual and its underlying assumptions is key.

# Looking ahead

1. Why Randomize? (June 15)
2. Ethics of Randomized Evaluations (June 22)
3. Generalizability (June 29)
4. Building Effective Academic-NGO Partnerships in the Humanitarian Space (July 13)

Thank you!

# References

Jones, Damon, David Molitor, and Julian Reif. "What do workplace wellness programs do? Evidence from the Illinois workplace wellness study." *The Quarterly Journal of Economics* 134.4 (2019): 1747-1791. https://doi.org/10.1093/qje/qjz023.

J-PAL Evaluation Summary: The Impact of a Workplace Wellness Program in Illinois

J-PAL Evaluation Summary: Workplace Wellness Programs to Improve Employee Health Behaviors in the United States

The Illinois Workplace Wellness Study Project Page

Pomeranz, Dina. "Impact evaluation methods in public economics: A brief introduction to randomized evaluations and comparison with other methods." Public Finance Review 45.1 (2017): 10-43.