J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

CASE 2: WORKPLACE WELLNESS PROGRAMS

Why Randomize?

Photo: Shutterstock.com

This case study is based on:

> 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 thanks the authors for allowing us to use their paper and for sharing their data.

LEARNING OBJECTIVES

To understand the different methods commonly used to estimate the impact of a given intervention, and to understand their strengths, weaknesses, and underlying assumptions.

SUBJECTS COVERED

Causality, counterfactual, impact, comparison groups, selection bias, omitted variables, randomization.

KEY VOCABULARY	
Comparison Group:	A group that is as similar as possible to the treatment group in order to be able to learn about the counterfactual. In an experimental design, the comparison group (also called the control group) is a group from the same population as the treatment group that, by random assignment, is not intended to receive the intervention.
Counterfactual:	What would have happened to program recipients had they not received the intervention. The counterfactual is a conceptual construct and is not something we can ever observe; it can only be inferred from a comparison group separate from the treatment group.
Estimate:	In statistics, a "best guess" about an unknown value in a population (such as the effect of a program on an outcome) based on a sample drawn from that population. For example, we may use average income in a random sample of citizens as our *estimate* of average income in the country as a whole.
Impact:	The impact of a program on a recipient is the difference between that individual's observed outcome given that they received the program and the counterfactual outcome that would have occurred had they not received the program. A program's impact on any one individual can never be observed; we can only estimate the *average impact* of a program across the entire sample by measuring the difference in the average outcome between the treatment and comparison groups.
Omitted Variable Bias:	Bias that occurs when important variables/characteristics are left out of the regression analysis. When these variables predict both the outcome and participation in an intervention, their omission can lead us to incorrectly over- or underestimate the impact of the program. For example, omitting the role of socioeconomic status, which is correlated with test scores, could lead to overestimating the impact of a tutoring intervention if wealthy students are also more likely to afford tutoring. In effect, the independent effect of wealth is misattributed to the tutoring program.
Treatment Group:	The group that receives the intervention.
Selection Bias:	Selection bias is bias that occurs when the individuals who receive the program are systematically different from those who do not. For example, consider an elective after-school tutoring program. Is it effective at raising children's exam scores? If we simply compare those who take up the tutoring program to those who don't, we will likely get a biased estimate of the effect of the tutoring program, because those who chose to participate are likely different from those who don't in terms of their academic performance (for example, those who took it up may be more motivated, or

	they may be weaker students). Randomization removes selection bias because it breaks the link between characteristics of the individual and their treatment status.

INTRODUCTION

How do we know if a program had an impact? This case study presents five different methods commonly used for estimating the impact of a policy or program, illustrating their strengths, weaknesses, and underlying assumptions. To motivate the concepts covered, we draw on a recent randomized evaluation of a workplace wellness program that was offered to faculty and staff at the University of Illinois at Urbana-Champaign.

Workplace wellness programs are one widely-touted solution to the rising cost of healthcare in the U.S. These programs seek to reduce medical spending by fostering healthy lifestyles among employees through the provision of free or subsidized health screenings, fitness programs, and classes on healthy behaviors like smoking cessation or stress management.

Estimating the impact of a workplace wellness program on employees' health and healthcare spending is difficult because it is impossible to know how healthy participants would have been had they never participated in the program, and by extension, how much money they (or their insurers) would have spent on healthcare. Ideally, evaluators would be able to track the health and healthcare spending of participants overtime as they participate in a program, measure any changes that occur, and then go back in time and measure the same group's progress without the program in place. This second, hypothetical set of outcomes represents what *would have happened* in the absence of treatment and is called the **counterfactual**.

Because we can never observe the counterfactual, the central challenge of any impact evaluation is to find a valid proxy for the counterfactual. We typically do this by selecting a group of people who resemble participants as much as possible but who did not participate in the intervention. This group is called the **comparison group**. It is important that the comparison group and the participant group are, on average, as similar as possible, so that we can attribute any differences in outcomes to the intervention. We can then estimate the **impact** by calculating the difference in outcomes observed at the end of the intervention between the comparison group and the **treatment group**.

A valid, unbiased impact estimate can only be attained if the comparison group is a good representation of the counterfactual. If the comparison group poorly represents the counterfactual, then the estimated impact will be **biased**, leading us to either over- or underestimate the true effect. The method used to select the comparison group is *the key decision* in the design of any impact evaluation. As we'll see in the case study,

Bias can result from a variety of factors that have the potential to make treatment and comparison groups different. **Selection bias** occurs when those who elect (or are *selected*) to participate in an intervention are different from those in the comparison group in terms of their pre-program outcomes (e.g., if healthier people are more likely to participate in workplace wellness programs). **Omitted variables bias** occurs when an external factor that determines your outcome also determines participation. Income, for example, could facilitate participation after-work wellness programs (through access to childcare) and health (through access to preventative medicine), leading to omitted variables bias if left out of the analysis.

The remainder of this case study implements each of the five methods using actual data from the Illinois Workplace Wellness Study. We illustrate the relative strengths, weaknesses, and underlying assumptions for each method, and we show how the different methods can produce very different results, leading to distinct and often conflicting conclusions about the efficacy of workplace wellness programs.

THE ILLINOIS WORKPLACE WELLNESS STUDY

Over the past several decades, healthcare costs in the U.S. have risen rapidly. According to the World Health Organization, healthcare costs accounted for over 17% of U.S. GDP in 2016, up from 12% in 2000 and more than any other country in the world.[1] Workplace wellness programs are one widely touted solution to the growing cost of healthcare in the U.S. These programs seek to reduce medical spending by fostering healthy lifestyles among employees, and often consist of activities like free health screenings, fitness programs, and classes on health-promoting behaviors such as smoking cessation, stress management, or nutrition.

Workplace wellness programs have become increasingly common. By 2018, more than 80 percent of large firms and more than half of small employers in the United States offered wellness programs, covering more 50 million workers.[2]

Yet despite the growing popularity of these programs, research on their effectiveness remains mixed, with studies finding negative, positive, or no impact on healthcare spending. However, because much of this literature relies on non-experimental methods that can sometimes be prone to bias, the true impact of these programs remains uncertain.

To address the need for credible evidence, researchers in J-PAL's network collaborated with the University of Illinois at Urbana Champaign to assess the impact of their workplace wellness program. The program, which was offered to a random sample of faculty and staff from September 2016 to September 2017, consisted of biometric health screenings,[3] a free health risk assessment,[4] and wellness activities such as fitness classes and seminars on healthy eating, smoking cessation, and other healthy habits.

The impact evaluation focused on three questions. First, did the program lead employees to live healthier, more active lifestyles? Second, did improvements in health lead to lower healthcare spending? And finally, did the program lead to fewer worker sick days and greater productivity, potentially paying for itself?

[1] World Health Organization Global Health Expenditure database (apps.who.int/nha/database).

[2] Kaiser Family Foundation, "Employer Health Benefits: 2016 Annual Survey," http://files. kff.org/attachment/Report-Employer-Health-Benefits-2016-Annual-Survey

[3] The biometric health screening included: height, weight, waist, and blood pressure measurements, a fingerstick prick to measure blood cholesterol, glucose, and triglycerides levels; and a consultation with a health coach to explain the measurements.

[4] The health risk assessment consisted of an online questionnaire designed to assess lifestyle habits.

With these questions in mind, the researchers collected data on:

- Healthcare spending from insurance claims data, including spending on in-patient and out-patient clinics and pharmaceutical spending
- Fitness habits, including visits to the campus fitness center per month, participation in the university's 10k race, and self-reported exercise habits
- Employee productivity, including job title / promotion, job retention, sick leave taken, hours worked per week, and self-reported job satisfaction and productivity
- Background variables such as age, gender, and socio-economic status

In the ensuing sections, we'll use these data to demonstrate commonly-used non-experimental approaches to impact evaluation, clarify the assumptions required for unbiasedness under each of these methods, and discuss when and whether these assumptions are likely to hold. Finally, we will benchmark these findings against the experimental results of the actual study.

Note: For clarity of exposition and to avoid directing focus away from the learning objectives, this case study omits discussion of statistical inference (e.g. confidence intervals and p-values).

APPLICATIONS TO OTHER CONTEXTS

The impact evaluation methods covered in this case study are applicable to any program where some people receive the intervention and some do not. Such 'selection' into program participation may occur either because a program is only provided to certain locations or target populations, or because some people elect to participate in an available program and others do not. Both forms of selection pose a challenge for evaluators, because they imply that program participants may be different from non-participants in ways that make comparisons between the two misleading.

The workplace wellness program in this study is a case where some people elect to participate and others do not. This situation also arises in the delivery of many social programs in low- and middle-income countries. For example, when evaluating the impact of microfinance loans on income, researchers have to consider that those who accept the offer of a loan may be more entrepreneurial than those who do not. Similarly, an evaluation of the impact of chlorine dispensers at water points on water quality and health would have to deal with the fact that those who purchase these items may be more health-conscious to begin with.

Ultimately, the goal of this case study is not to learn about workplace wellness programs, but to understand the methods available for evaluating a program's impact. Accordingly, when reading this case study, consider how the methods and takeaways from this study translate to social programs in your area of work.

ESTIMATING THE IMPACT OF THE ILLINOIS WORKPLACE WELLNESS PROGRAM

METHOD 1: SIMPLE DIFFERENCE BETWEEN PARTICIPANTS AND NON-PARTICIPANTS

Workplace Wellness Programs Lead to More Active Lifestyles and Lower Healthcare Costs: (Fictitious) Evidence from the Illinois Workplace Wellness Study

Abstract

Objective: To assess the impact of workplace wellness programming on employees' fitness habits, health, healthcare spending, and workplace productivity.

Sample*:* 3,300 employees (staff and faculty) at the University of Illinois at Urbana-Campaign.

Intervention*:* Biometric screening (height, weight, BMI, and blood pressure), health risk assessment and wellness activities such as fitness classes and lessons on smoking cessation and stress management.

Outcomes*:* Visits per month to campus fitness center; Average monthly spending on healthcare; index of workplace productivity based on employee retention, sick leave, promotion, and job satisfaction.

Design*:* Comparison of average outcomes between employees participating in at least one program activity and employees not participating in any program activity.

Results*:* On average, participants visited the gym nearly twice as often as non-participants (7.3 times per month versus 3.8) and spent 25% less on healthcare ($650 per month versus $500). Despite these differences, they were only marginally more productive at work.

Conclusion*:* Workplace wellness programs are a promising avenue for promoting employee fitness and reducing healthcare spending, but they do not appear to increase worker productivity.

Key Figures

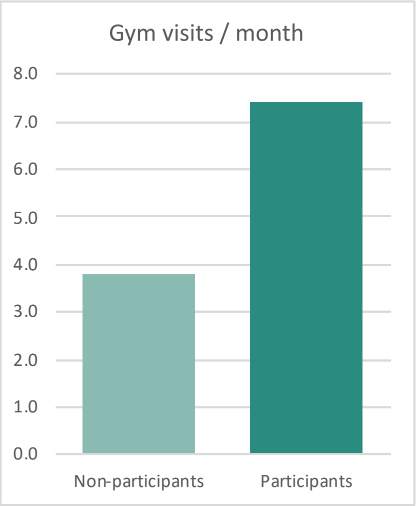

Figure 1a

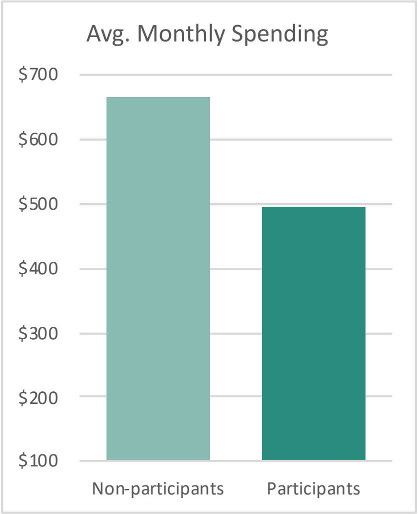

Figure 1b

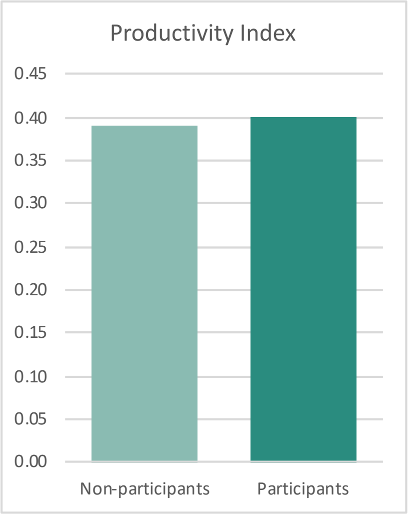

Figure 1c

DISCUSSION

1. What is the comparison group in this study? What is the counterfactual?

2. What assumptions are necessary for this counterfactual to be valid? To what extent do you think these assumptions are likely to hold in the context of the Illinois Workplace Wellness study? What might be some potential violations?

METHOD 2: PRE VS. POST COMPARISON

Workplace Wellness Programs Increase Spending, Decrease Worker Productivity, and Fail to Improve Fitness: (Fictitious) Evidence from the Illinois Workplace Wellness Study

Abstract

Objective: To assess the impact of workplace wellness programming on employees' fitness habits, health, healthcare spending, and workplace productivity.
***Sample**:* 3,300 employees (staff and faculty) at the University of Illinois at Urbana-Campaign.
***Intervention**:* Biometric screening (height, weight, BMI, and blood pressure), health risk assessment and wellness activities such as fitness classes and lessons on smoking cessation and stress management.
***Outcomes**:* Visits per month to campus fitness center; Average monthly spending on healthcare; index of workplace productivity based on employee retention, sick leave, promotion, and job satisfaction.
***Design**:* Longitudinal pre-post study, comparing participants' outcomes before versus after the intervention was delivered.
***Results**:* During the course of the program, participants' gym visits declined by 4%, healthcare spending increased by 21%, and workplace productivity declined by 26%.
***Conclusion**:* Workplace wellness programs fail to improve fitness and appear to have adverse effects on healthcare spending and worker productivity, possibly because they encourage the overuse of unnecessary care, pushing spending higher and drawing workers away from productive work.

Key Figures

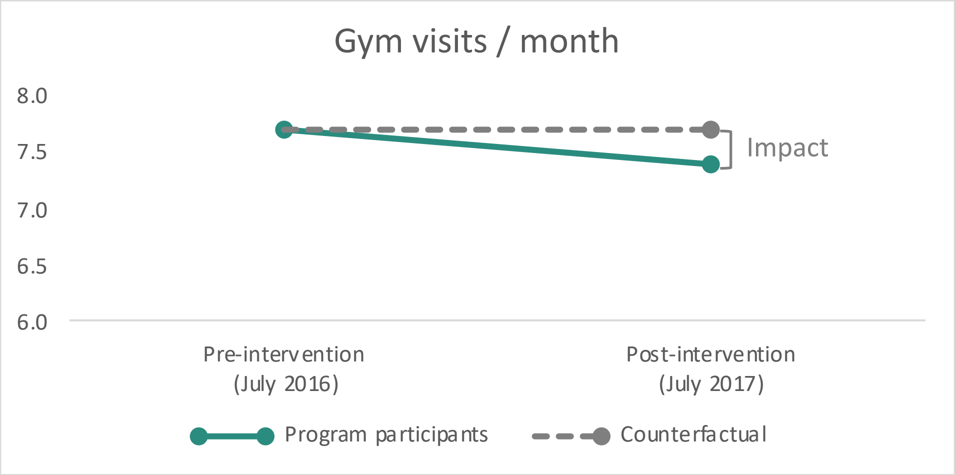

Figure 2a

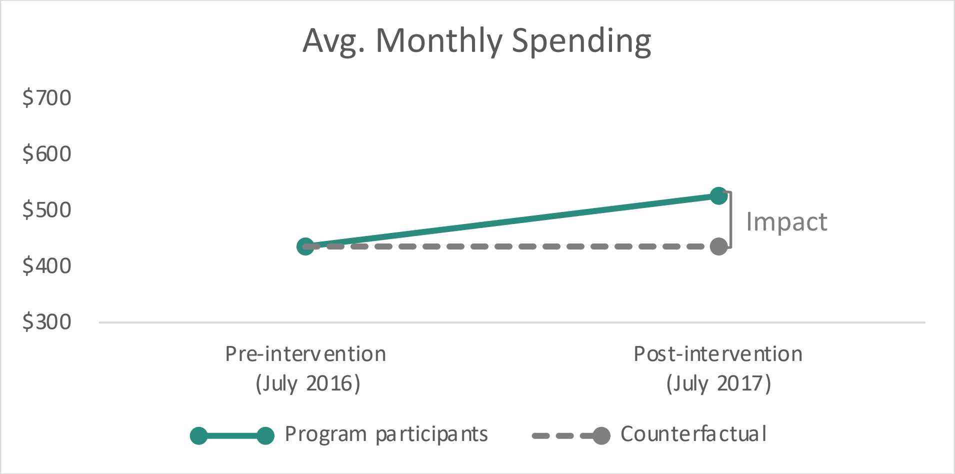

Figure 2b

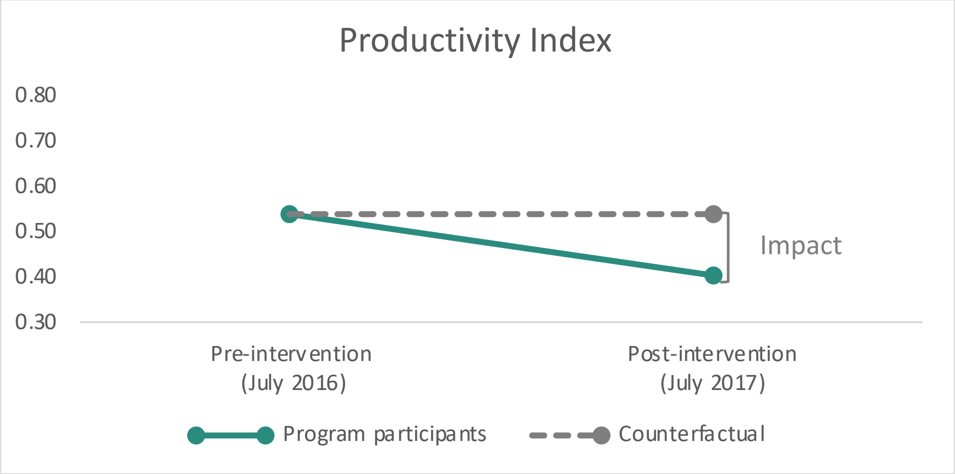

Figure 2c

DISCUSSION

1. What is the comparison group in this study? What is the counterfactual?

2. What assumptions are necessary for this counterfactual to be valid? To what extent do you think these assumptions are likely to hold in the context of the Illinois Workplace Wellness study? What might be some potential violations?

METHOD 3: DIFFERENCE-IN-DIFFERENCES

Workplace Wellness Programs Improve Fitness but Fail to Reduce Spending or Improve Workplace Productivity: (Fictitious) Evidence from the Illinois Workplace Wellness Study

Abstract

Objective: To assess the impact of workplace wellness programming on employees' fitness habits, health, healthcare spending, and workplace productivity.

Sample*:* 3,300 employees (staff and faculty) at the University of Illinois at Urbana-Campaign.

Intervention*:* Biometric screening (height, weight, BMI, and blood pressure), health risk assessment and wellness activities such as fitness classes and lessons on smoking cessation and stress management.

Outcomes*:* Visits per month to campus fitness center; Average monthly spending on healthcare; index of workplace productivity based on employee retention, sick leave, promotion, and job satisfaction.

Design*:* A difference-in-differences research design is used to compare changes over time among program participants to changes overtime among non-participants, with the *difference* in these *differences* taken as the estimate of impact.

Results*:* Using a difference-in-differences design to simultaneously account for over time changes and selection bias, we find that the Workplace Wellness Program increased visits to fitness facilities by 18% relative to pre-intervention levels, but had small and insignificant effects on healthcare spending (-5%) and workplace productivity (4%).

Conclusion*:* Workplace wellness programs can increase fitness, but these changes are unlikely to be large enough to lead to meaningful improvements in health, reductions in healthcare spending, or greater workplace productivity.

SUMMARY OF THE DIFFERENCE-IN-DIFFERENCES DESIGN

The *difference-in-differences* design combines the simple-difference and the pre-vs-post comparison designs to simultaneously account for selection bias and over time trends. The basic idea behind this method is to compare the difference between participants and non-participants before the program to the difference between them after the program, taking the *difference* in these overtime *differences* as the impact estimate, as depicted in Figure 3a. If outcomes improve more for participants than for non-participants between pre- and post- intervention periods, then the difference in the differences would be positive, and we would conclude the program had a positive impact.

Because this design compares participants to themselves overtime, it is robust to selection bias. And because it differences-out overtime changes among non-participants left unaffected by the program, the design accounts for external forces or overtime trends that affect both groups equally, such as country-level trends in healthcare costs.

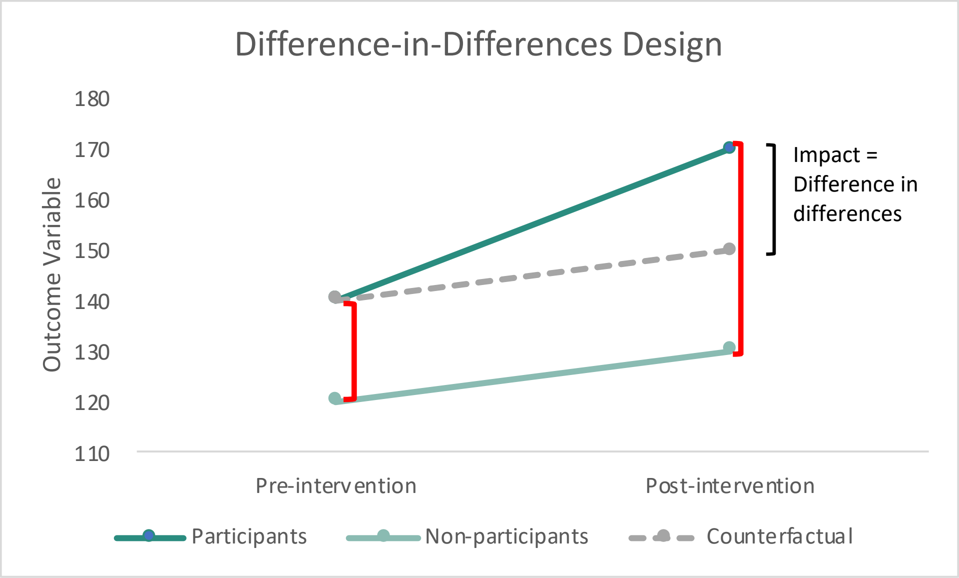

Figure 3a

Figures 3b-3d display the estimated impact on of the Illinois Workplace Wellness program using the difference-in-differences design:

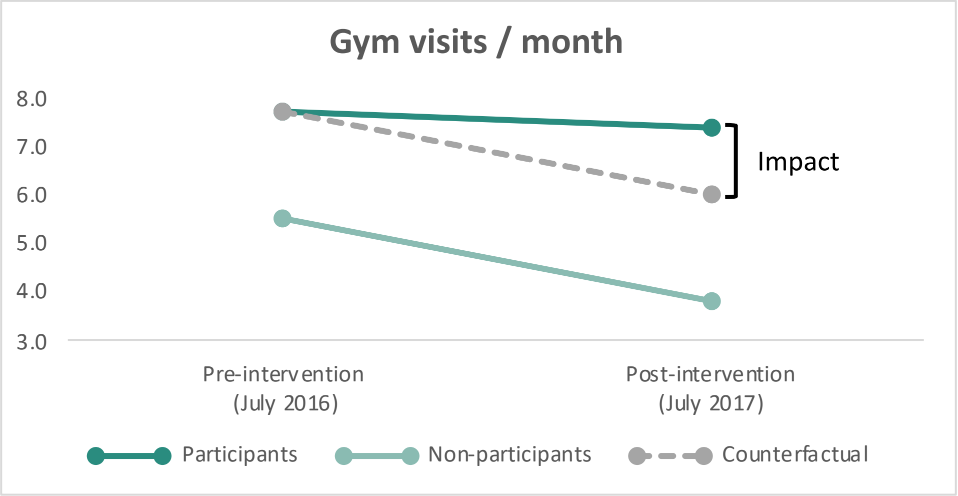

Figure 3b

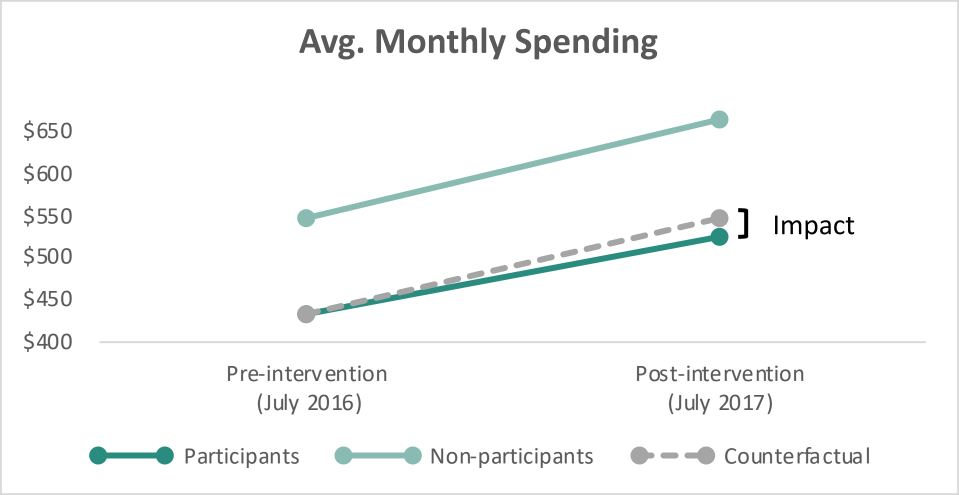

Figure 3c

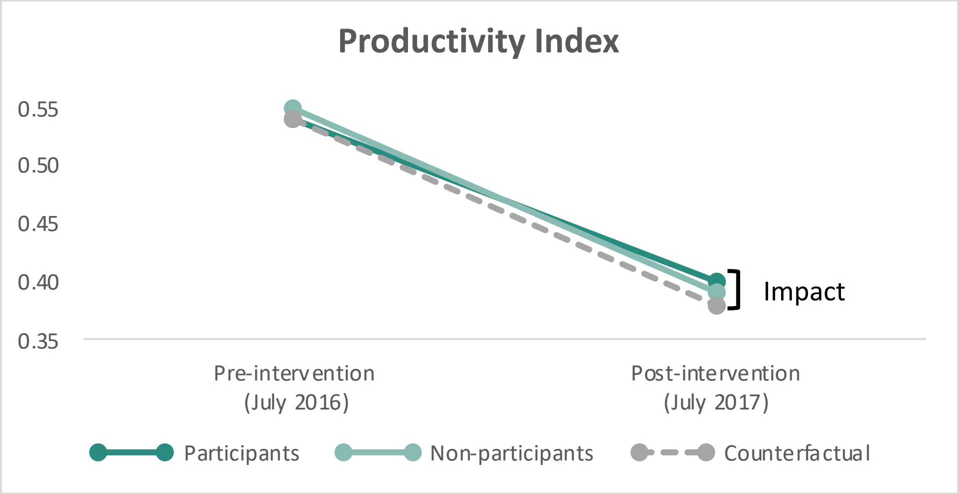

Figure 3d

DISCUSSION

1. What is the comparison group in this study? What is the counterfactual?

2. What assumptions are necessary for this counterfactual to be valid? To what extent do you think these assumptions are likely to hold in the context of the Illinois Workplace Wellness study? What might be some potential violations?

METHOD 4: MATCHING AND REGRESSION

Workplace Wellness Programs Reduce Healthcare Spending but Fail to Improve Fitness or Productivity: (Fictitious) Evidence from the Illinois Workplace Wellness Study

Abstract

Objective: To assess the impact of workplace wellness programming on employees' fitness habits, health, healthcare spending, and workplace productivity.

Sample*:* 3,300 employees (staff and faculty) at the University of Illinois at Urbana-Campaign.

Intervention*:* Biometric screening (height, weight, BMI, and blood pressure), health risk assessment and wellness activities such as fitness classes and lessons on smoking cessation and stress management.

Outcomes*:* Visits per month to campus fitness center; Average monthly spending on healthcare; index of workplace productivity based on employee retention, sick leave, promotion, and job satisfaction.

Design*:* Statistical control strategies (OLS regression and matching) are used to compare employees who participated in the program to employees that did not participate in the program but are otherwise similar in terms of their age, gender, race, income, and other measured variables that might account for health and healthcare spending.

Results*:* The workplace wellness program failed to significantly increase visits to fitness facilities or work place productivity, but it did reduce healthcare spending by roughly $146 per month (28%).

Conclusion*:* Workplace wellness programs reduce healthcare spending, but not through healthier fitness habits. Downstream impacts on workplace productivity are limited, casting doubt on whether this approach is worthwhile for employers.

SUMMARY OF MATCHING AND REGRESSION RESEARCH DESIGNS

As we discussed in previous sections, differences between participants and non-participants in terms of age, gender, socio-economic status, and other characteristics imply that non-participants are unlikely to be a valid counterfactual for participants, confounding both the simple-difference and difference-in-differences designs.

Table 1 takes a closer look at these differences in the case of the Illinois Workplace Wellness Study, comparing those who elected to participate in the wellness program to those who did not. It shows that participants and non-participants differed significantly on the basis of gender, salary, and likelihood of being a faculty member. Perhaps most importantly, they varied on the basis of their pre-intervention levels of the study's key outcome variables --- healthcare spending, frequency of gym visits, and workplace productivity.

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
Productivity index (pre-intervention)	0.55	0.54	0.01*	3251

Male	46%	40%	6%**	3300
Age	44.1	43.6	0.4	3300
Above median 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.

When researchers only have cross-sectional data, the two most common approaches to handle such confounding factors are *matching* and *regression*.

Matching

The idea behind matching is to minimize confounding factors by constructing a comparison group that is as similar as possible to the treatment group in terms of their observable characteristics. In the simplest form of matching, each participant is matched to a non-participant with identical characteristics. These matched individuals then become the comparison group, and the remaining unmatched individuals are excluded from the analysis (Figure 4a).

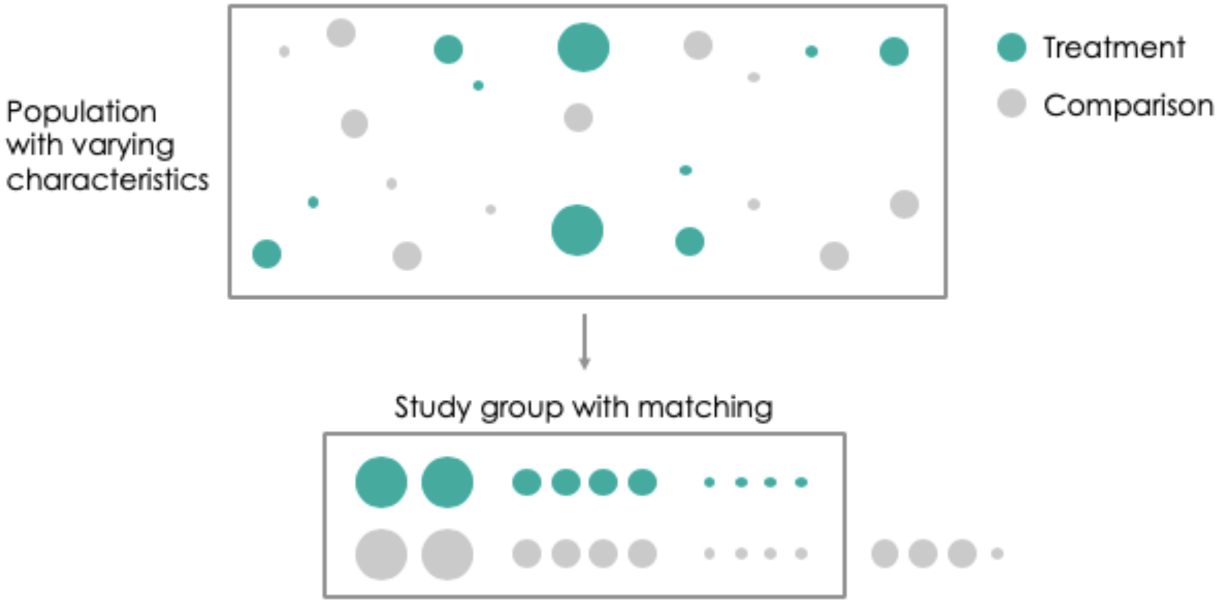
Population with varying characteristics
Treatment
Comparison
Study group with matching

Figure 4a

“Exact matching” only works when the number of matching variables is small (or the dataset, and by extension the pool of potential matches, is very large). Matching algorithms become more complicated as the number of matching variables increases, but the basic idea remains the same – treated individuals are matched to untreated individuals who are otherwise as similar as possible in terms of their pre-treatment characteristics, although not identical.

Table 2 shows the results of a matching algorithm applied to the Illinois Workplace Wellness Study. Whereas without matching, participants spent an average of $103 less per month before the intervention than non-participants, in the matched sample this difference drops to a modest and statistically insignificant $19.1. Differences on other variables also decrease, suggesting that non-participants are more comparable to participants in the matched sample than in the full sample and thus more likely to be a valid counterfactual. However, improvements in comparability come at the cost of a reduced sample size, from $N = 3300$ in the full sample to $N = 1109$ in the matched sample (participants who had no comparable non-participant match (and vice versa) were dropped). Moreover, while the matched sample is well-balanced on these "observable" variables – variables that were measured in the pre-intervention survey and included in the matching algorithm – there is no guarantee that they will be balanced on "unobservable" variables that were not measured in the survey or variables that were measured but not included in the matching algorithm. For instance, it could be that participants and non-participants look very different in terms of their eating habits, a difficult-to-measure yet potentially important determinant of healthcare spending.

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
Productivity index (pre-intervention)	0.54	0.54	0.0	1109
Male	0.45	0.39	6%**	1109
Age	43.1	42.7	0.4	1109
Above median salary	48%	48%	0%	1109
Faculty	11%	8%	3%	1109

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

Assuming that non-participants in the matched sample are indeed a valid counterfactual for participants, we can move forward with estimating the program's impact on post-intervention fitness habits, healthcare spending, and workplace productivity:

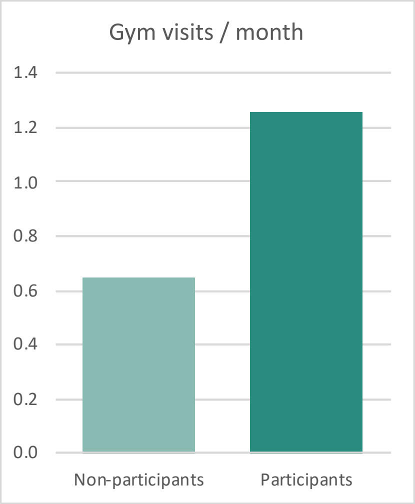

Figure 4a

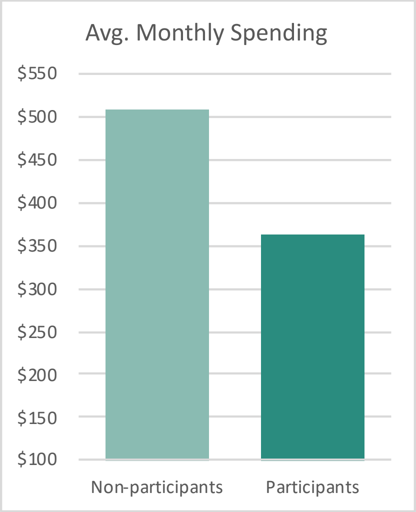

Figure 4b

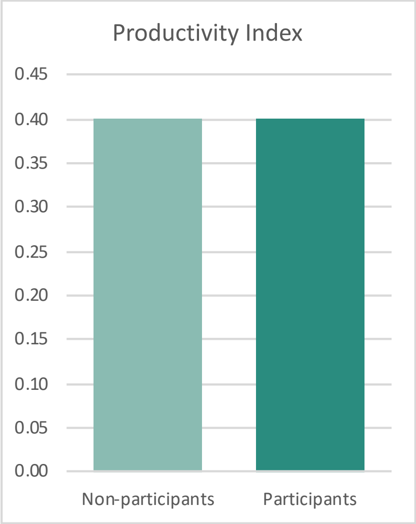

Figure 4c

Regression (Optional)

Regression is similar to matching in that it is a method of accounting for pre-intervention differences between participants and non-participants in order to compare "like with like." However, whereas matching accounts for these differences by trimming the sample to eliminate non-participants with no comparable participant match (and vice versa), regression accounts for differences by *modeling* the relationship between the outcome of interest and the set of potentially confounding variables, then comparing the predicted outcomes across participants and non-participants. If (and only if!) this model is correctly specified and includes all of the potentially confounding variables --- two very strong assumptions --- then the results of the model will have a causal interpretation.

For instance, in the current example, we might specify the following regression model to account for confounding factors:

$$Spending_i = \beta_0 + \beta_1 Participates_i + \beta_2 Male_i + \beta_3 Age_i + \beta_4 White_i + \epsilon_i$$

Where $Spending_i$ is our outcome for individuals index by i, $Participates_i$ is an indicator for participants in the program (equal to 1 for participants and 0 for non-participants), ϵ_i is an error capturing the difference between predicted spending and actual spending, and the rest of the control variables follow from Table 1, above. We would estimate the parameters of this model --- β_0 through β_4 --- using data from the full sample. If we were confident the model was correctly specified and included all relevant variables, we would interpret β_1 as the impact of participation in workplace wellness programs on healthcare spending.

DISCUSSION

1. What is the comparison group in this study? What is the counterfactual?

2. What assumptions are necessary for this counterfactual to be valid? To what extent do you think these assumptions are likely to hold in the context of the Illinois Workplace Wellness Study? What might be some potential violations?

3. Optional: To what extent do the results from this method differ from those of Method 1 (Simple Difference)? What might account for these differences?

METHOD 5 - RANDOMIZED EVALUATION

Recognizing the potential pitfalls of non-experimental methods, researchers in J-PAL's network conducted a randomized evaluation to experimentally test the impact of the workplace wellness program. After enrolling 4,834 employees in the study, 3,300 were randomly assigned to have access to the program, and 1,534 were assigned to a comparison group. Of those in the treatment group with access to the program, 56 percent (1,848) participated by completing the health screening and health risk assessment. Even though not everyone in the treatment group participated, this relatively high rate of uptake still allowed the researchers to compare average levels of healthcare spending, fitness, and workplace productivity across treatment and control groups with enough statistical power to detect relatively small effects.[5]

Figures 5a-5c depict these comparisons graphically, showing small and insignificant differences across all three study outcomes.

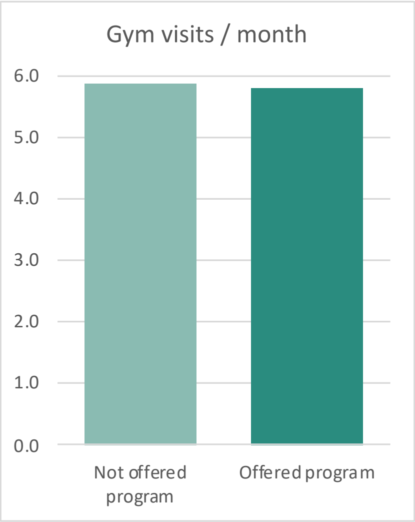

Figure 5a

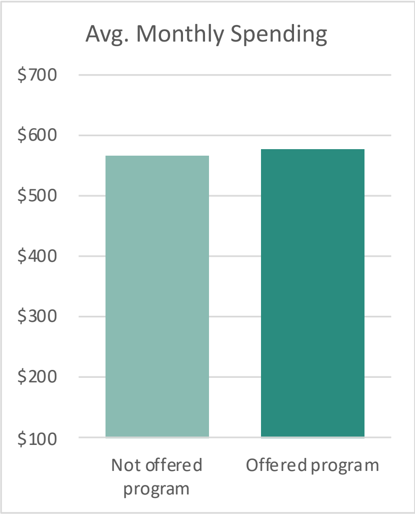

Figure 5b

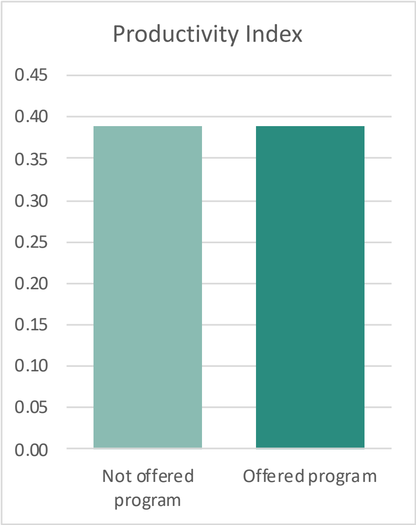

Figure 5c

[5] In this study, we focus on the "Intent to Treat" (ITT) effect, which is the effect of *offering* workplace wellness programs. The ITT should not be confused with the effect of actually participating in the program for those who opt to do so. Estimating this quantity is more involved and beyond the scope of this case study.

DISCUSSION

1. What is the comparison group in this study? What is the counterfactual?

2. What assumptions are necessary for this counterfactual to be valid? To what extent do you think these assumptions are likely to hold in the context of the Illinois Workplace Wellness Study? What might be some potential violations?

COMPARING ALL FIVE METHODS:

Table 3 presents the impact estimates of the workplace wellness program using the five different methods discussed in this case study.

Table 3: Comparison of impact estimates across methods

Method	**Gym visits / month**	**Healthcare Spending**	**Productivity Index**
Simple difference	3.5**	-$137**	0.01
Pre-post	-0.4	$100**	-.15**
Difference-in-differences	1.34**	-$9.6	0.01
Matching	0.61	-$146**	0.00
Randomized evaluation	-0.06	$10	0.00

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

As is apparent from the table, not all methods yield the same result, and many of the non-experimental methods yield estimates that do not match the experimental benchmark. Moreover, in the cases where a non-experimental estimate does happen to align with the experimental estimate, it does so for only one or two of the three outcomes. These discrepancies between the experimental and non-experimental results imply that non-experimental evaluations of workplace wellness programs are misleading due to selection bias and/or unmeasured differences between participants and non-participants.

Generalizing beyond the case of workplace wellness, randomized evaluations are the most credible method to estimate impact because they are the only method that can decisively rule out selection bias and unmeasured confounding. When randomized evaluations are not practical or feasible, non-experimental methods should be used with caution and with a clear understanding of the counterfactual and the plausibility of the underlying assumptions.

REFERENCES AND FURTHER READING

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.

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.

REUSE AND CITATIONS

To request permission to reuse this case study or access the accompanying teachers' guide, please email training@povertyactionlab.org. Please do not reuse without permission. To reference this case study, please cite as:

> J-PAL. "Case Study: The Illinois Workplace Wellness Study: Why Randomize?" Abdul Latif Jameel Poverty Action Lab. 2020. Cambridge, MA.

	Method	Description	What assumptions are required, and how demanding are the assumptions?	Required data
Randomization	Randomized Evaluation/ Randomized Control Trial	Measure the differences in outcomes between randomly assigned program participants and non-participants after the program took effect.	*The outcome variable is only affected by program participation itself, not by assignment to participate in the program or by participation in the randomized evaluation itself*. Examples for such confounding effects could be information effects, spillovers, or experimenter effects. As with other methods, the sample size needs to be large enough so that the two groups are statistically comparable; the difference being that the sample size is chosen as part of the research design.	Outcome data for randomly assigned participants and non-participants (the treatment and control groups).
Basic Non-Experimental Comparison Methods	Pre-Post	Measure the differences in outcomes for program participants before the program and after the program took effect.	*There are no other factors (including outside events, a drive to change by the participants themselves, altered economic conditions, etc.) that changed the measured outcome for participants over time besides the program*. In stable, static environments and over short time horizons, the assumption might hold, but it is not possible to verify that. Generally, a diff-in-diff or RDD design is preferred (see below).	Data on outcomes of interest for program participants before program start and after the program took effect.
	Simple Difference	Measure the differences in outcomes between program participants after the program took effect and another group who did not participate in the program.	*There are no differences in the outcomes of participants and non-participants except for program participation*, and both groups were equally likely to enter the program before it started. This is a demanding assumption. Non-participants may not fulfill the eligibility criteria, live in a different location, or simply see less value in the program (self-selection). Any such factors may be associated with differences in outcomes independent of program participation. Generally, a diff-in-diff or RDD design is preferred (see below).	Outcome data for program participants as well as another group of non-participants after the program took effect.
	Differences in Differences	Measure the differences in outcomes for program participants before and after the program *relative* to non-participants.	*Any other factors that may have affected the measured outcome over time are the same for participants and non-participants, so they would have had the same time trajectory absent the program*. Over short time horizons and with reasonably similar groups, this assumption may be plausible. A "placebo test" can also compare the time trends in the two groups before the program took place. However, as with "simple difference," many factors that are associated with program participation may also be associated with outcome changes over time. For example, a person who expects a large improvement in the near future may not join the program (self-selection).	Data on outcomes of interest for program participants as well as another group of non-participants before program start and after the program took effect.

	Method	Description	What assumptions are required, and how demanding are the assumptions?	Required data
More advanced statistical non-experimental methods	Multivariate Regression/OLS	The "simple difference" approach can be—and in practice almost always is—carried out using multivariate regression. Doing so allows accounting for other observable factors that might also affect the outcome, often called "control variables" or "covariates." The regression filters out the effects of these covariates and measures differences in outcomes between participants and non-participants while holding the effect of the covariates constant.	Besides the effects of the control variables, *there are no other differences between participants and non-participants that affect the measured outcome*. This means that any unobservable or unmeasured factors that do affect the outcome must be the same for participants and non-participants. In addition, the control variables cannot in any way themselves be affected by the program. While the addition of covariates can alleviate some concerns with taking simple differences, limited available data in practice and unobservable factors mean that the method has similar issues as simple difference (e.g., self-selection).	Outcome data for program participants as well as another group of non-participants, as well as "control variables" for both groups.
	Statistical Matching	Exact matching: participants are matched to non-participants who are identical based on "matching variables" to measure differences in outcomes. Propensity score matching uses the control variables to predict a person's likelihood to participate and uses this predicted likelihood as the matching variable.	Similar to multivariable regression: *there are no differences between participants and non-participants with the same matching variables that affect the measured outcome*. Unobservable differences are the main concern in exact matching. In propensity score matching, two individuals with the same score may be very different even along observable dimensions. Thus, the assumptions that need to hold in order to draw valid conclusions are quite demanding.	Outcome data for program participants as well as another group of non-participants, as well as "matching variables" for both groups.
	Regression Discontinuity Design (RDD)	In an RDD design, eligibility to participate is determined by a cutoff value in some order or ranking, such as income level. Participants on one side of the cutoff are compared to non-participants on the other side, and the eligibility criterion is included as a control variable (see above).	*Any difference between individuals below and above the cutoff (participants and non-participants) vanishes closer and closer to the cutoff point.* A carefully considered regression discontinuity design can be effective. The design uses the "random" element that is introduced when two individuals who are similar to each other according to their ordering end up on different sides of the cutoff point. The design accounts for the continual differences between them using control variables. The assumption that these individuals are similar to each other can be tested with observables in the data. However, the design limits the comparability of participants further away from the cutoff.	Outcome data for program participants and non-participants, as well as the "ordering variable" (also called "forcing variable").
	Instrumental Variables	The design uses an "instrumental variable" that is a predictor for program participation. The method then compares individuals according to their predicted participation, rather than actual participation.	*The instrumental variable has no direct effect on the outcome variable. Its only effect is through an individual's participation in the program*. A valid instrumental variable design requires an instrument that has no relationship with the outcome variable. The challenge is that most factors that affect participation in a program for otherwise similar individuals are also in some way directly related to the outcome variable. With more than one instrument, the assumption can be tested.	Outcome data for program participants and non-participants, as well as an "instrumental variable".