Tool 4.2 – Did it Work? Interpreting Study Findings

Data analytics projects often seek to answer the question of what works best, or did an intervention or strategy improve a desired outcome. These efforts can lead project teams to make causal claims. For example, if the analysis shows that program participants who attended job club were employed at a faster rate than those program participants who did not attend job club, you might be tempted to claim that the job club improved employment outcomes. But before making such a claim, it is important to interrogate your analytics results to make sure that they can support your causal claim. For example, it is not enough to simply compare those who received the job club to those who did not receive the job club and attribute the difference to the job club itself. Part of interrogating your results is assessing whether there are plausible alternative explanations for your findings. Any plausible alternative explanations may threaten your analysis's internal validity, which is the extent to which you can be confident that the cause-and-effect relationship claimed in your study does not have alternative explanations. If internal validity is weak or threatened, it means that you can't make a causal claim about whether your intervention achieved the desired outcome.

This tool is designed to assist in interpreting findings from three common nonexperimental research designs for estimating program outcomes: pre-/post-test, interrupted time series (ITS), and comparative interrupted time series (CITS) designs. These designs are nonexperimental designs since they do not involve random assignment and are typically used when there is data on outcomes over time. This tool does not suggest when these research designs are appropriate to use or how to use but offers basic information on each.

In most cases, these research designs are not set up in a way that will allow you to make a causal claim. Simply looking at outcomes before and after an intervention is rarely enough to demonstrate a cause/effect relationship. However, the designs described below have improved in recent years and smart implementation of design features can greatly improve the rigor. Lastly, these designs are presented in order of least to most complexity and rigor with pre-/post-test being the least to CITS being the most.

Instructions: Outlined below is a brief definition of the study design, a sample graph and a checklist to use to assess internal validity for each type of research design followed by a sample graph of findings.

Pre-/Post-Test Design

Pre-/post-test designs involve comparing outcomes for individuals at two different points in time—before and after an intervention or policy change is implemented—to see if there is a change in the outcome.[1] For example, programs may compare the average earnings of their

1. Thiese (2014).

participants prior to entering the program to their average earnings after they complete the program. This can be helpful to describe the outcomes of individuals in a program (for example, the average earnings of individuals who attend the program), but it does not provide information on the effectiveness of a program at improving an outcome.[2]

Sample Graph

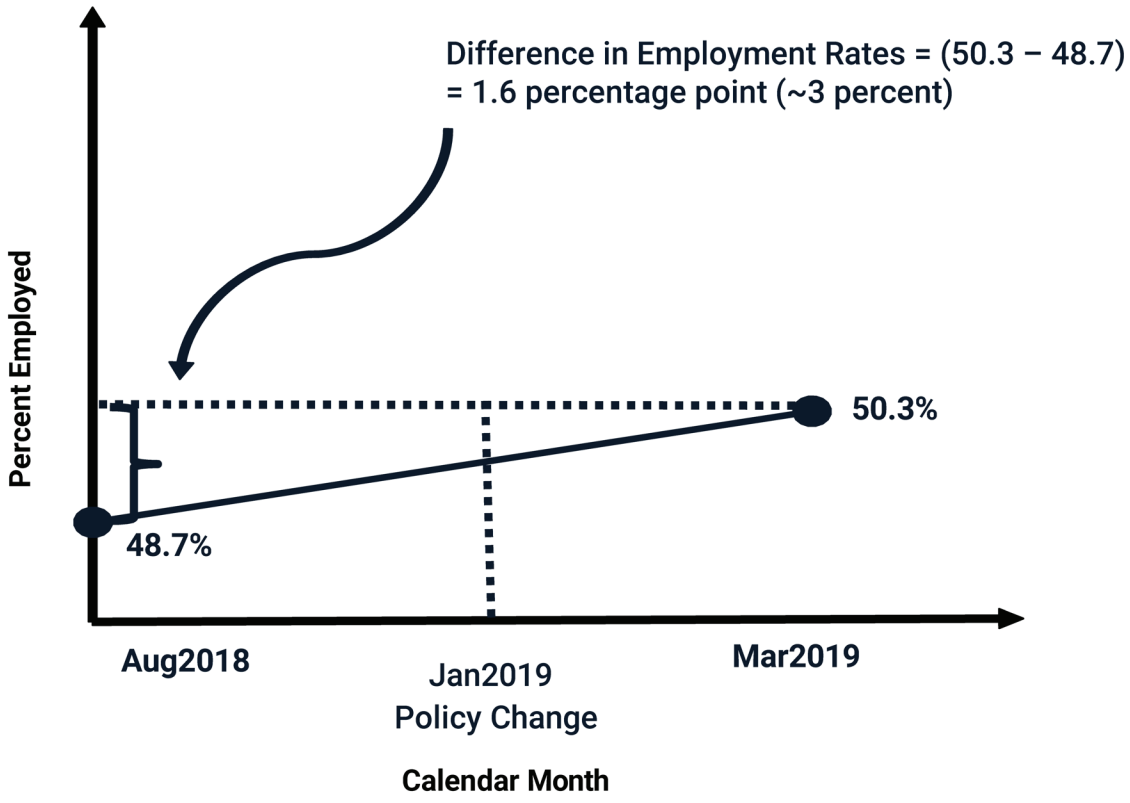
In order to increase employment among TANF recipients, the State of Sufficia made two changes to its CCDF-childcare funded services in 2019:

1. Prioritized TANF clients for childcare services

2. Waived co-payments for TANF families

To evaluate these program changes, the State of Sufficia sought to compare the average employment rate of TANF recipients before and after the two changes were made in 2019. Sufficia found a 1.6 percentage point increase in the employment rate between August 2018 and March 2019, as illustrated in Figure 4.2.1. It should be noted that this type of design does not allow Sufficia to say that the policy changes led to this increase in employment, as it does not account for other potential reasons the employment rate might have increased.

Figure 4.2.1 Employment Rate Before and After Sufficia Policy Changes



2. Gopalan, Rosinger and Bin Ahn (2014).

Validity Checklist

There are several questions to ask when considering or implementing a pre-/post-test design. The answers to these questions can help inform whether you should consider doing a pre-/post-design, and help you better understand the results you see and the claims you can make about them. As shown in Table 4.2.1, focus on whether these potential threats are *plausible*, and not just *possible*.

Table 4.2.1 Pre-/Post-Test Design: Validity Checklist

VALIDITY THREAT	PLAUSIBLE?	ADDRESSED?
History: Are other events occurring concurrently (for example, a new employer opened in the local area)?		
Maturation: Is there naturally occurring change in the outcome over time (for example, people's earnings tend to increase as they accumulate work experience)?		
Selection: Do characteristics of who is in the sample used to measure the outcome vary by time point?		
Attrition: Are study subjects dropping out? Are you missing outcome data for some people?		
Testing: Are behavior changes occurring due to the act of measuring the outcome (for example, staff are more aware, so they are entering data more consistently)?		
Instrumentation: Did the outcome measure/way the outcome is collected change over time?		

Interrupted Time Series (ITS)

Interrupted time series research designs involve collecting data on outcomes at multiple time points, both before and after an intervention or policy change is implemented, to assess whether there is a difference in the "trend" of the outcome after the implementation of the intervention or policy change (compared to before). Or in other words, whether the intervention or policy change "interrupts" the existing trend in the outcome.[3]

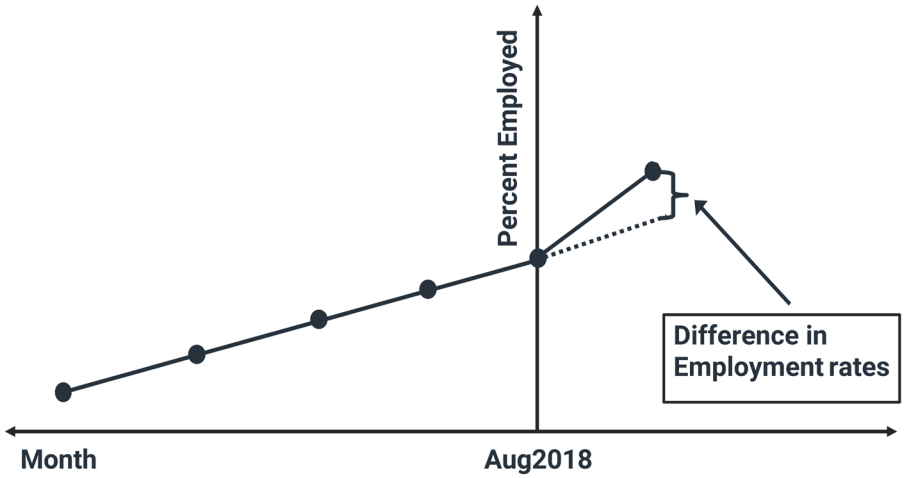
This design is more rigorous than a pre-/post-test design, as having outcome data in multiple time periods can help you see changes that are occurring outside of the intervention or policy change. However, it still does not always allow you to say with confidence that a program led to an effect on an outcome, as it does not control for other outside factors that could also influence the outcome.

3. St. Clair, Hallberg & Cook (2014).

Sample Graph

The state of Sufficia collected data on employment every month in the year before and in the month after implementing the two program changes mentioned above. The monthly employment rates are shown in Figure 4.2.2. The vertical line shows when the program changes were implemented (in August 2018).

Figure 4.2.2 Employment Rate at Multiple Points Before and After Policy Change



Sufficia discovered that the employment rate observed following the program changes deviated from the employment rate trend in the months prior to the changes; the increase was greater than the projected increase in employment. This is seen by the actual employment rate being higher than the line showing the projected employment rate. It should be noted that this type of design does not allow Sufficia to say with full confidence that the program changes led to this increase in employment, unless they have accounted for all other potential reasons the employment rate might have increased.

Validity Checklist

There are several questions to ask when considering or implementing an ITS design. In general, ask yourself: Does the ITS study address these potential alternative explanations for changes in outcomes? As shown in table 4.2.2, focus on whether these potential threats are *plausible*, and not just *possible*. For ITS studies, look for abrupt *changes* in these aspects *at the time of the intervention*.

Table 4.2.2 Interrupted Time Series: Validity Checklist

VALIDITY THREAT	PLAUSIBLE?	ADDRESSED?
History: Are other events occurring concurrently (for example, a new employer opened in the local area)?		
Instrumentation: Did the outcome measure/way the outcome is collected change over time?		
Special issue with multiple cross-sectional time series		
Selection: Do characteristics of who is in the sample used to measure the outcome vary by time point?		
Special issue with panel/longitudinal time series		
Maturation: Is there naturally occurring change in the outcome over time (for example, people's earnings tend to increase as they accumulate work experience)?		
Attrition: Are study subjects dropping out? Are you missing outcome data for some people?		
Testing: Are behavior changes occurring due to the act of measuring the outcome (for example, staff are more aware, so they are entering data more consistently)?		

Comparative Interrupted Time Series (CITS)

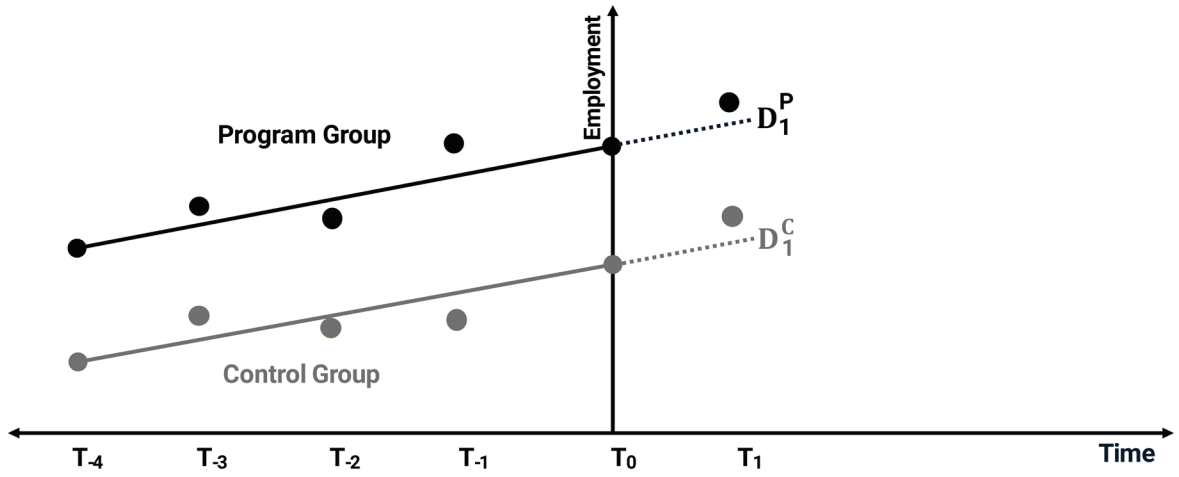
Similar to the ITS design, the CITS design involves collecting outcome data at multiple time points before an intervention and during at least one time point after an intervention. The difference is in this design, outcome data is also collected for a group of individuals that did not experience the intervention or policy at the same time. This inclusion of a "control" or "comparison" group helps you see what would have happened in the absence of the intervention, and whether any outside changes affected the outcomes for those individuals. The CITS design allows for a more robust evaluation.[4]

Sample Graph

Sufficia decided to compare the employment rates of two groups of participants: one that was subject to the policy changes described above and one that was not. Figure 4.2.3 illustrates the employment outcomes for the program group (the group that was subject to the policy changes, shown in black) and for the control group (the group that did not receive the intervention, shown in gray) at multiple time points before the policy changes and at one point after the policy changes.

4. St. Clair, Hallberg, and Cook (2014).

Figure 4.2.3 Estimated Effect of the TANF Policy Changes on the Employment Rate



Note: DIP and DIC are the deviation from the baseline trend for the program group and control group, respectively.

Sufficia was able to estimate the effect of the intervention by taking the difference in the deviation from the employment rate trends between the two groups after the intervention (in time = T_1).

Validity Checklist

There are several questions to ask when considering or implementing the CITS design. In general, ask yourself: does the CITS study address these potential alternative explanations for any effects that it finds? As shown in Table 4.2.3, focus on whether these potential threats are *plausible*, and not just *possible*.

Table 4.2.3 Comparative Interrupted Time Series: Validity Checklist

VALIDITY THREAT	PLAUSIBLE?	ADDRESSED?
Selection: Are the average characteristics of individuals in the program and control groups different?		
Do any of the following affect one research group *differently than the other*?		
History: Do the events occurring concurrently with the intervention differ (for example, another policy change made by a different agency)?		
Maturation: Does the naturally occurring change in the outcome over time differ (for example, people's earnings tending to increase as they work more)?		
Selection: Do characteristics of who is in the sample used to measure the outcome vary by time point?		
Regression artifacts: That is, the tendency of extreme scores to gravitate toward the mean.		
Attrition: Does the rate of study subjects dropping out differ? Does the rate of missing outcome data differ?		
Instrumentation: Do any changes in the way the outcome is measured/collected differ?		