The Benefits of Using Longitudinal Data when Conducting a Randomized Control Trial

Paul Glewwe
University of Minnesota

"Covid-19 and Longitudinal Research: Opportunities and Challenges"

May 12, 2021

Introduction

Randomized control trials (RCTs) are increasingly being used to assess the impact of programs in both developing and developing countries.

In the simplest case, RCTs can be implemented using cross-sectional data collected after the program has been in place "long enough": ***Panel (longitudinal) data are not necessary*.**

However, there are several benefits of collecting panel data when implementing and RCT.

In this presentation, I present what I see as the main advantages of collecting panel data when conducting an RCT. I divide them into 4 types:

- To check for "threats to identification" (threats that could lead to bias)
- To increase efficiency (both statistical and budgetary)
- To improve "endline" (post-program) data collection
- To provide more "sophisticated" estimates of program impacts

I. Check for Threats to Identification

"In theory", RCTs provide unbiased/consistent estimates of the impact of the program. **Yet *many things can go wrong with RCTs***.

Two common problems are:

1. Lack of balance between the control group and the treatment group(s)

2. Sample attrition before the endline (post-treatment) data are collected

Panel data, and in particular ***collection of baseline data***, can be used both to test for the problem and to remedy the problem.

Lack of Balance between Control Group and the Treatment Group(s)

Random assignment of individuals (or groups, or communities) to the control group and the treatment group(s) should lead to groups that are almost identical, but **sometimes random chance leads to situations where the groups are not balanced**.

Balance needs to be checked *before* the program is implemented, since the program could cause changes in many (though not all) variables. ***It is particularly important to check whether the outcome variable of interest is balanced between the treatment and control group(s).***

If there is imbalance in certain variables, their baseline values should be entered into the regression equation to control for this lack of balance.

Sample Attrition

It is often the case that ***some individuals*** who are in the (hopefully balanced) control and treatment groups when the evaluation starts ***cannot be found, or refuse to participate, when the endline*** (post-treatment) ***data are collected***.

If this attrition is random, this is not a problem, but attrition could be non-random.

Collecting baseline data allow for calculation of attrition rates and patterns, e.g. to see whether attrition is random or correlated with observed variables.

If attrition appears to be non-random, there methods that can be used to estimate upper and lower "bounds" on the treatment effects, such as the method of Lee (2009). Yet these methods typically require panel data to implement them.

II. Increase Efficiency (Statistical and/or Budgetary)

Even if there are no problems that lead to bias, collecting only one round of data after the program has been in place "long enough" may not be the most efficient way to estimate the impact of the program.

High Correlation between Baseline and Endline of Outcome Variable

If the outcome variable is **highly correlated over time** (correlation of baseline value with endline value, denoted by ρ), then the **standard error of the estimated treatment effect** (denoted by SE($\hat{\beta}$)) will be **smaller when panel data are collected** (σ^2 is variance of outcome variable):

$$SE(\hat{\beta})_{\text{endline only}} = \sqrt{2\sigma^2/n} \quad (n = \text{number in each group})$$

$$SE(\hat{\beta})_{\text{baseline and endline}} = \sqrt{\left(\frac{2\sigma^2}{n}\right) 2(1-\rho)}$$

So if $\rho > 0.5$ then $2(1-\rho) < 1$ and $SE(\hat{\beta})_{\text{baseline and endline}} < SE(\hat{\beta})_{\text{endline only}}$

Baseline May Provide "Valid" Control Variables to Increase Precision

The standard errors just presented are based on regressions with no other variables in them, but adding "control" variables to a regression may increase precision of the estimates.

But the ***control variables are "valid" only if they are not affected by the program***.

By definition, baseline variables cannot be affected by the program since the program has not yet started.

Thus baseline data can provide control variables that may increase the precision of the estimated treatment effects.

Multiple Rounds of Endline Data Can Increase Precision of Estimates

Some outcome variables may be very "noisy" because they are hard to measure accurately. One example of this is the income or profits of household businesses.

If this measurement error has little or no correlation over time, a "less noisy" estimate of the outcome variable will be to collect two or more endline (post-treatment) rounds of that variable. This is done by averaging the outcome variable over these rounds.

III. Improve Endline Data Collection

Another benefit of collecting baseline data is that allows the team to improve the quality and/or usefulness of the endline data. Here are two examples of this:

Learning by Doing

In practice, collecting data from households or other "units" is very messy, and the ***quality of the data*** collected ***increases with experience***.

Collecting baseline data typically reveals information that can be used to improve the quality of the data collection "instruments" (questionnaires, health measurements, tests of academic skills), and more generally the quality of the data collection "system".

The lessons learned from baseline data collection can be used to improve the collection of the endline data.

Baseline Data Can be Used to Revise the "Pre-Analysis Plan"

It is becoming more common among social science researchers, when conducting RCTs, to write (and "publish") a "Pre-Analysis Plan". This is done before the endline data are collected, and even before the baseline data are collected. It is done to "tie the hands" of researchers so that they do not "mine the data" to find some "significant" impact of the program.

Writing the Pre-Analysis Plan before any data are collected is difficult because little may be known about the context, for example about the variation in the variables of interest.

In many cases it may be "acceptable" to revise the Pre-Analysis Plan after doing some exploratory analysis with the baseline data (but **before** endline data are collected).

Example: The baseline data should reveal what control variables are closely correlated with the outcome of interest.

IV. Provide More Sophisticated Estimates of Impact

A final set of advantages for collecting panel data when conducting an RCT is that such data allow one to produce more "sophisticated" estimates of the impact of the program. Here are three examples of this.

Provide Useful Variables for Estimating Heterogeneity of Impacts

Programs are unlikely to have the same impact on all members of the population of interest, so it is useful to estimate impacts separately for groups of particular interest.

Some of the variables that defined groups could be affected by the program, so it is best to collect data on those variables at baseline.

Example: The impact of an education program could vary by the initial learning levels of students. By collecting data on learning levels at baseline, one can see whether the program works best for weaker students.

Check for Longer Run Impacts of the Program

The impacts of any program could last for many years, or they could "fade out" quickly. More generally, one would like to estimate how program impacts evolve over time.

This can be done by collecting endline (post-program) data at two or more points in time, such as immediately after participants have "graduated" from the program, 1-2 years later, and 3-5 years later. Indeed, the outcomes measured could change.

Example: For an education program at the primary school level, collect test score data from students when they have finished primary school, and when they have finished other levels of schooling. Also, collect wage or income data when they have reached "working age".

Check "Speed" at which Effects Are Generated

For many programs, the "treatment" can take place over different amounts of time. It is possible that most or all of the effect of a program can take place relatively quickly, so that extending the treatment for a longer period of time produces little benefit.

To see when a program's benefits occur, it is useful to collect "midline" data, that is data on the outcome of interest (and possibly other variables) before the program is finished. This is done in addition to collecting endline (post-treatment) data.

Example: An early childhood program that provides "coaching" to mothers on how to care for their children's health and educational development. How many months of coaching are needed until the women acquire most or all of the skills that the program is designed to teach?

Thank you!

Questions?

Comments?