Check for updates

CHAPTER 3

Available Testing Techniques

Abstract This chapter discusses some necessary definitions and the pros and cons of the techniques available for field experiments of the Universal Basic Income. These techniques include randomized controlled trials, saturation studies, and combinations of the two.

Keywords Basic income experiments • Negative Income Tax experiments • Social science experiments • Basic income • Universal Basic Income • Randomized controlled trial • Saturation study • Saturation studies

After this chapter defines some relevant terms, it discusses the pros and cons of the techniques available for testing UBI.

All empirical research (whether experimental or not) attempts to answer a question appropriately called the research question. Often a large study, like a UBI experiment, will ask a series of research questions. A question like "what are UBI's effects" is too vague to be useful. A UBI could have an infinite number of effects, some important and some trivial. Although researchers would be happy to discover effects they were not looking for, you can't find an effect that you make no effort to measure.

© The Author(s) 2018

K. Widerquist, *A Critical Analysis of Basic Income Experiments for Researchers, Policymakers, and Citizens*, Exploring the Basic Income Guarantee, https://doi.org/10.1007/978-3-030-03849-6_3

Most research questions are formulated around hypothesis testing. That is, they test a claim about a supposed relationship. For example, a lot of medical research tests the hypothesis that a medical treatment is safe.

Empirical studies seldom conclusively verify or falsify a claim. They can only state whether the evidence is consistent with or contradictory toward the claim, but this much is often extremely useful.

Sometimes there is little doubt that a treatment has a particular effect, and the research question becomes, "How large is that effect?" That sort of a research question is useful to examine, but to be a hypothesis test, it has to be paired with the claim that the effect is larger, smaller, or equal to some amount. For example, in wealthy countries, past evidence indicates UBI will correspond with a decline in the average time recipients spend in employment. The question is: how much it will decline? What size of a finding would be significant? Is it that the response is greater than zero? If so, we don't need a test. Is it that the response is greater than X%? If so, among which group? Is it that it is large enough to make the program unsustainable? Or is it something else entirely: perhaps the significance of this response is not in how large it is, but in some qualitative measure of what people do with the reduced time they spend working? The differences between these potential research questions create problems discussed in later chapters.

Two desirable attributes for estimates are that they are "accurate" and "unbiased." An accurate estimate is one that is likely to be close to the actual value. An "unbiased" estimate is one that is just as likely to overestimate the actual value as it is to underestimate it. That is, it lacks "statistical bias." Statistical bias is very different from the bias in the sense of favorability to one group over another.

Statistical bias cannot always be eliminated, and sometimes it has to be traded off against accuracy. A slightly biased estimation technique could be preferable to an unbiased but less accurate measure. For example, suppose you were estimating a person's age. A biased technique is likely to produce results anywhere from 1 year below to 2 years above their actual age. An unbiased technique is likely to produce an estimate anywhere from 20 years below to 20 years above their actual age. The accuracy of the biased technique almost certainly makes it more useful.

Bias causes great difficulty for empirical studies. Sometimes you don't know whether a technique is biased or not. Sometimes you know that it is likely to be biased, but you don't know which way. Sometimes you know that it is likely to be biased in a particular direction, but you don't know how much. All of these problems affect the testing of UBI.

One surprisingly controversial definitional issue is what to call the effort to try out UBI on a small scale to learn something about it in advance of full implementation. In common English, the words "test," "trial," "pilot," and "experiment" all fit that definition, but some of them are also used in more specific senses in technical settings.

"Experiment" is sometimes used to refer only to a "randomized controlled trial" (RCT): a test designed to isolate the effects of the factors being studied by using randomization as a method to control as much as possible for all other factors that might influence the relevant outcomes. Researchers do so by randomly selecting two sufficiently large groups that differ as little as possible from each other and from the wider population. They give the treatment to one group only (the experimental group) and observe whether that group differs in relevant ways from the other group (the control group). If the groups are sufficiently large and properly selected, the differences between them—other than those caused by the treatment—will tend to cancel each other out. This method is indispensable in many forms of medical research, and it can be useful in social science as well. But as argued below, it is not always the best way to address questions at issue in the UBI debate.

"Pilot" or "pilot project" can be used as a broader alternative to "experiment," but it carries baggage as well. "Pilot project" sometimes implies that the test is conducted by an authority with the power to fully implement the policy—at least if the pilot meets some criteria of success. Sometimes it implies that a firm decision in favor of full implementation has already been made, and the test is being used to determine *how* rather than whether to implement it.

Even the simple word "test" sometimes implies that the study involves some firm criteria by which the policy will be judged to have passed or failed. Nonspecialists often expect such criteria from experiments of any kind. Social science experiments are usually conducted without any criteria of success in mind in a context where success criteria are politically controversial debates. Therefore, it's best to fight the impression they have any such criteria.

The term "trial" or "implementation trial" has the fewest other connotations, and so I occasionally use it for clarity, but it is also the least familiar of these terms.

I mostly use the term "experiment" in that broader sense defined in the first paragraph of this chapter, despite how, as explained below, at least some specialists assert the common usage is wrong.

What distinguishes an experiment, test, trial, or pilot in this broad sense from a nonexperiment is that an experiment is in place solely (or at least primarily) to learn something about a potential policy. It is not (primarily) an attempt to *implement* the policy. In this sense, the NIT experiments of the 1970s were experiments, but the Alaska Dividend and Cherokee per capita payments, for example, are not.[1] Although these policies might provide a useful opportunity to learn something about UBI, they are not put in place for that opportunity.

Once the decision is made to conduct an experiment, researchers have a choice of two broad types of techniques or a combination of the two. The first, an RCT, is defined above. The second, a "saturation study," involves identifying two relevant communities, such as two small towns, and giving the treatment to everyone in one community and not to people in the other. Although researchers might randomly choose which of the two sites will be the control and which the experimental site, that level of randomness is not enough to control for other factors that might make one site different from another. Although the communities could be selected to be as similar to each other and to the wider population in as many observed ways as possible, they might differ in important but unobserved ways.

Both RCTs and saturation studies are useful. RCTs are better at examining issues in which most of the effects occur at the individual level. Though far from perfect, saturation studies are better at examining issues in which many important effects occur at the community level. These "community effects" are extremely important for UBI because its effects depend on the interactions of people in markets and cultural settings (see discussion below).

Whether the trial is an RCT or a saturation study, the experimental and control groups each need to be at least a few hundred (and preferably a few thousand people) to produce statistically useful results. How large the sample has to be depends on "the law of large numbers," a statistical principle stating that as the number of observations increases in an unbiased sample, the probability of the expected accuracy of that sample increases.

[1] Karl Widerquist and Michael W. Howard, eds., *Alaska's Permanent Fund Dividend: Examining Its Suitability as a Model* (New York: Palgrave Macmillan, 2012); *Exporting the Alaska Model: Adapting the Permanent Fund Dividend for Reform around the World* (New York: Palgrave Macmillan, 2012).

The law of large numbers begins to kick in between 20 and 30 observations, and for most purposes, 50 observations is enough to provide a high likelihood that the results should be highly accurate.

That makes UBI experiments sound affordable, but suppose you want results for men and women. Now you need 100 observations. Suppose you need a statistically useful sample of children, and people of various ethnic and religious groups. Now you need several hundred observations. Suppose you want to observe the effects of UBI on unemployment or pregnancy. Now you need well into the thousands, so that the number of people who become pregnant or employed during the study is statistically significant. Although a UBI experiment with a few hundred participants can produce useful results for some issues, most experiments usually try to get funding for a sample well into the thousands to examine more issues.

In wealthier countries, a sample of a few thousand people receiving a meaningfully large UBI is extremely expensive. But in less wealthy countries, where people live off extremely small incomes, much larger sample sizes are possible—perhaps into the tens of thousands. Thus, doing different kinds of experiments in different places is extremely useful.

Once the experimental group is selected and begins receiving "the treatment," researchers observe how they behave in comparison to the control group. The central goal of any experiment is to find a way to ensure differences between the two groups will be attributable as much as possible to the treatment and to random fluctuations, which tend to cancel out in a large enough sample. Hence the *control* in the experiment. Unfortunately, in social science, creating a trial that is both controlled and representative of how the policy under investigation will work under full implementation is extremely difficult.

Some researchers—labeled "Randomistas" by their critics—insist that only RCTs are truly scientific or truly deserving of the term "experiment."[2] One reason to resist the Randomista use of "experiment" is to avoid confusion caused by the belief that more technical definitions are the "right" definitions. That is not how language works. Specialists do not own the language or any terms within it. The most commonly used definition is the most acceptable definition. Specialists who insist that technical definitions

[2] Guy Standing, "Basic Income Pilot Schemes: Seventeen Design and Evaluation Imperatives," in *Wege Zum Grundeinkommen [Pathways to Basic Income]*, ed. D. Jacobi and W. Strengmann-Kuhn (Berlin: Bildungswerk Berlin, 2012).

are the only right definitions risk confusing nonspecialists, who are most familiar with the common understanding of "experiment" and who are important consumers of the findings of UBI experiments—or any policy-related experiment.

Another reason to resist the Randomista use of the word is that RCTs are not accurately described as the only scientific form of experiment.[3] RCTs make some valuable statistical techniques available that aren't available with saturation studies, and they make it possible to control for unobserved factors that saturation studies cannot control for. But they do so by entirely ignoring certain kinds of effects (discussed below). In other words, RCTs control for more things but test fewer things. Therefore, researchers should be open to using both RCTs and saturation studies as appropriate. Both techniques should be considered part of the social scientists' toolkit as long as researchers are careful to note the extent to which their results should be seen as tentative or conclusive and the ways in which those results are likely to be biased.

Each technique has some advantages over the other in each of these respects: important effects of UBI occur at both the individual and the community level. Individuals immediately react to UBI in many important ways that are worth estimating, but they interact with other individuals in markets, society, culture, and politics. All of these interactions generate important feedback effects throughout the community. Existing theory and empirical evidence indicate that some community effects might be as important or more important than the initial individual effects of UBI. If researchers opt only for an RCT they must choose between ignoring feedback effects entirely or supplementing their experimental data with information from other sources to simulate feedback. Guy Standing argues that the Randomista attitude often leads to ignoring community effects even on issues—such as UBI—where such effects are likely to be extremely important.[4]

Because both types of experiments have advantages and disadvantages, an ideal test would fully combine saturation and RCT techniques by randomly selecting dozens of saturation sites for both the control and experimental groups. For example, consider a test of whether a vaccine creates "herd immunity," which refers to the way a large number of individuals

[3] Andrew Gelman, "Experimental Reasoning in Social Science," in *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, ed. Dawn Langan Teele (New Haven, CT: Yale University Press, 2014).

[4] Standing.

with immunity in a group helps protect individuals without it. The individual immunity question can be answered by a simple RCT with a few hundred or a few thousand individual subjects, but the herd immunity question requires testing multiple herds. The effort becomes more difficult if we need to test how large or isolated the herd must be to establish herd immunity. These questions might require dozens or even hundreds of herds of varying sizes and levels of isolation to get statistically significant results. For herds of livestock, such a test might be affordable. For herds of humans, it is likely probably unaffordable.

Researchers have conducted experiments with multiple saturation sites in India and Kenya, where poverty is extremely high and a UBI of a dollar a day or less is extremely significant to recipients. The Kenyan study has the budget for a statistically significant number of saturation sites, but each site is too small to capture all of the relevant community effects, many of which probably occur at the national level.

Most likely, in wealthier countries, the techniques available will be limited to one RCT or one saturation site, or at best one of each.