CHAPTER 4

Testing Difficulties

Abstract This chapter discusses several general problems that virtually any experiment in the Universal Basic Income will have to deal with: community effects, long-term effects, the Hawthorne effect, the streetlight effect, and the difficulty of separating the effects of the size and type of program being studied.

Keywords Basic income experiments • Negative Income Tax experiments • Social science experiments • Basic income • Universal Basic Income • Community effects • Feedback effects • Hawthorne effect • Streetlight effect

This chapter discusses several difficulties that are likely to affect any UBI experiment and possible ways of dealing with each one, including community effects, the Hawthorne effect, the streetlight effect, and the difficulty of separating the effects of the size and type of policy being studied.

1 Community Effects

Community effects (defined in Chap. 3) will probably have a large impact on many, if not most, of the responses to UBI. This section explains why these effects create enormous difficulties for UBI experiments and makes some tentative suggestions about how to deal with them.

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_4

Community effects are easiest to grasp when they work in the same direction as individual effects. For example, evidence indicates that inequality and the ghettoization of poverty exacerbate problems like ill-health, crime, poor education, and so forth, and sometimes inequality makes these problems worse even for the people who materially benefit from inequality.[1] If an individualized RCT finds that UBI has a positive effect on childhood health at the individual level, we can imagine that the effect will be even larger at the national level.

Community effects are more difficult to grasp when they (fully or partially) counteract individual effects. In such cases, the national effect might be much smaller or even the reverse of the more easily observable individual effects. For example, some obvious and important community effects of UBI have to do with the feedback effects between workers and employers, most particularly the labor demand response. Workers (at least in wealthier nations) are likely to respond to UBI by working less. Employers are likely to respond to that action by offering better wages and working conditions. Workers are likely to respond to better wages and working conditions by working more, partially counteracting their initial drop in hours worked. Call that a feedback loop. It involves the supply and demand for labor and for related goods. Many researchers have criticized RCTs—and all field experiments—for their inability to examine general equilibrium effects,[2] which are important not just to wages, working conditions, and working hours, but to all economic variables.

Culture, education, and other factors are likely to respond to those changes in the labor market, and these factors could feedback to other labor market changes. That feedback loop now has five potential steps. An RCT can measure only the first step in the six steps in that predicted loop. A saturation study might capture some of the second and third steps, but only to the extent that these effects occur at the local level. Therefore, an experiment will tell us very little about what we want to know about hours worked, wages, and the incomes of workers. All of these factors will have an important effect on the cost of UBI.

[1] Richard G. Wilkinson and Kate Pickett, *The Spirit Level: Why More Equal Societies Almost Always Do Better* (London: Allen Lane, 2009).

[2] Angus Deaton, "Instruments, Randomization, and Learning About Development," 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), p. 177; Philippe Van Parijs and Yannick Vanderborght, *Basic Income: A Radical Proposal for a Free Society and a Sane Economy* (Harvard University Press, 2017), p. 143.

Ideally, the extent to which feedback loops cause these effects is something we would like to investigate in an experiment. To do so, we would need a prohibitively expensive version of the herd immunity test described in Chap. 3. Many the of relevant community effects will be observable only at the national level, but a saturation study might pick up enough of them to be useful.

Researchers with limited budgets have at least four options for dealing with community effects. Each of them has a serious downside. First, conduct an RCT only and ignore community effects entirely: concentrate on explaining the difference in behavior between the control and experimental groups without concern for (an accurate) national prediction. This option, clearly the worst of the four, biases the results, sometimes in unpredictable ways, and even if the direction of bias is predictable, the size of the bias seldom is.

Second, conduct an RCT only, leaving all the biases in place, but include caveats explaining those biases. This option is likely to be popular with researchers, but it has many shortcomings. Specialists often have difficulty explaining caveats in ways nonspecialists can understand in the time they have. Readers often ignore them because they are usually tedious and difficult to understand. Caveats often get lost in the chain of communication connecting specialists to citizens and policymakers. In practice, this second option might not be that different from the first. The 1970s US experiments attempted this option, but as Chap. 6 shows, the public discussion proceeded with little or no recognition that unobservable community effects existed.

Third, conduct an RCT in combination with computer simulation analysis using theory and data from other sources to estimate community effects. This option means the report on the experimental findings will be driven less by those findings and more by the assumptions of that simulation model. Hopefully, the assumptions of those simulation models will be drawn from very good evidence, but evidence to the quality we want is seldom available.

Fourth, conduct a saturation study on at least one site (more if budget allows), combined (if budget allows) with an individualized RCT at another site or across a wide geographical area. Small, isolated communities are likely to have community effects more similar to those we can expect at the national level. For example, if the saturation site is fairly isolated, local businesses have to draw labor from potential employees who are all eligible for UBI rather than from nearby neighborhoods that are not

involved in the study. Unfortunately, labor markets, even in isolated communities, are in many ways national and so even a saturation study is likely to be biased toward underestimating employer response, but they are an improvement on RCTs, which are unable to estimate employer responses at all. A saturation study won't provide evidence about how similar the community effects at the saturation site are to the community effects of a national program. Additionally, individuals in smaller, more isolated communities might not be representative of the people in larger, less isolated communities, where the majority of the world's population lives. This imperfect representativeness will bias the study in unknown ways.

2 The Hawthorne Effect

The "Hawthorne effect" is the problem of people changing their behavior when being observed. People in an experiment know they're being observed, and this knowledge might affect their behavior in unpredictable ways, causing many different forms of bias. Perhaps seeking approval of the observers, participants would behave in ways they think will make them look good or smart or successful to the observers. Perhaps instead they would show off, trying to be funny or interesting or trying to cultivate some kind of image. Perhaps they would try to "help" the observer by displaying what they think the observer wants to see. Perhaps they would try to "harm" the observer by displaying the opposite of what they think the observer wants to see, possibly because of some antagonistic feelings toward either the researcher or the research objective. Perhaps they would be affected by the power of suggestion: knowing that the observer wants to know whether they do X might unconsciously make them do X or make them avoid doing X more than they normally would. These reactions might sound silly, but no one can claim to be completely free of them. Hawthorne effects have been recognized for decades, but exactly how they are likely to affect research remains a mystery,[3] making it very difficult to compensate for them. One strategy is to observe people in an unobtrusive way for a long period of time in hopes that they gradually stop paying attention to their observers, but this strategy's success rate is hard to gage.

Hawthorne effects are likely to be a bigger problem for the new round of UBI experiments than they were in the 1970s. Today, most people post

[3] Jim McCambridge, John Witton, and Diana R. Elbourne, "Systematic Review of the Hawthorne Effect: New Concepts Are Needed to Study Research Participation Effects," *Journal of Clinical Epidemiology* 67, no. 3 (2014).

about themselves on social media, and it will be difficult to get them to avoid posting about a trial they are participating in. This visibility will make it easier for the media to find them, and the more attention they receive for participating in a study, the greater the Hawthorne effect is likely to be.

Saturation studies are more vulnerable to the Hawthorne effect than RCTs. A saturation site cannot be kept secret. Participants might have journalists, bloggers, activists, and long-lost friends contacting them to ask what it's like to be in the UBI saturation study.[4] How this increased attention will affect their behavior is unknown. I hope the problem does not make it impossible to do saturation studies in well-wired countries, but it might.

3 Long-Term Effects

Any experiment is going to be short term compared to how long the actual policy is likely to stay in place, and short-term effects often differ significantly from long-term effects. This problem is intuitively easy to grasp for people with no special training, but its magnitude is so great that it might create problems for understanding research. In most cases, the experimental UBI will be in place for only 2–4 years, while an actual UBI will be in place permanently, and we most want to understand its final, overall, long-term effects.

The effects of UBI on health, education, labor time, wages, working conditions, and so on are likely to involve community effects that develop out of economic and cultural interactions between people over a very long period. Experiments directly observe only the initial steps in that long, complex chain of reactions. Although some long-term effects are likely (at least) to be in the same direction as short-term effects, other long-term effects might partially or fully reverse the short-term effects. Following up with participants 5, 10, or 20 years after a temporary study has been completed is useful to see whether it has had lingering effects, but the lingering effects of a temporary policy are very different from the long-term effects of a policy that continues in place for 20 or more years. For example, some evidence indicates that the British labor force took as long as 70 or 80 years to react fully to the introduction of that nation's pension system.[5]

[4] Thanks to Evelyn Forget for alerting me to this last issue.

[5] Paul Johnson, "Parallel Histories of Retirement in Modern Britain," in *Old Age from Antiquity to Post-Modernity*, ed. Paul Johnson and Pat Thane (London: Routledge, 1998); http://blog.spicker.uk/experiments-with-basic-income-were-never-going-to-settle-the-arguments/

Researchers can try running a longer-term experiment, but doing so increases the expense and the time it takes to get results, and so most studies are very short term. The Seattle/Denver Income Maintenance Experiment (SIME/DIME) study contained the longest-run observations so far. It was originally planned for 6 years. After about 3 years, researchers obtained permission to extend the experiment to 20 years for a small subsample, but that effort was cancelled after 9 years.[6] That is, a small group was eligible for an NIT for 9 years, about six of which they were led to believe they would receive the NIT for 20 years. Researchers did not find major differences between this group and the shorter-term sample, but this RCT had no way to measure community effects, which are likely to be larger in the long run. How differently a national UBI would affect people over the long term still remains questionable. The best we can do is to extrapolate based on theory and data from other sources, imposing yet more assumptions about things we would rather like to learn from an experiment.

4 The Streetlight Effect

Although the "streetlight effect" is easy to understand, it might be the most difficult problem for experiments to avoid.

The streetlight effect gets its name from a joke in which a man loses his keys in a dark alley but looks for them under a streetlight because, he explains, "the light is so much better here." In social science, the "streetlight effect" is research that focuses on questions that are easier to answer but less important rather than on questions that are more important but harder to answer.

Few, if any, research techniques can examine all questions we have about a policy. Any study using any one technique draws attention to the questions that technique is better able to address and distracts attention from other, possibly more important questions.[7]

[6] P.K. Robins, "The Labor Supply Response of Twenty-Year Families in the Denver Income Maintenance Experiment," *Review of Economics and Statistics* 66, no. 3 (1984); Widerquist, "A Failure to Communicate: What (If Anything) Can We Learn from the Negative Income Tax Experiments?"

[7] Dawn Langan Teele, "Introduction," in *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, ed. Dawn Langan Teele (New

A social science experiment is a tool to help evaluate a potential policy. What's ultimately important about an experiment is its ability to do that. But an experiment is also a very specific tool that is much better at addressing some questions than others. Even the questions it can address, it can address only partially and/or indirectly—thereby producing information that is substantively different and possibly distracting from the most important information for the evaluation of that policy.

Experiments will find useful evidence, but understanding its value requires remaining focused on the big, evaluative questions and making the difficult, sometimes tenuous connection between that evidence and the important questions.

But research reports, academic literature, and popular literature on past experiments have overwhelmingly focused on the things experiments were best able to observe—differences between the control and experimental groups—as if those differences were the most important issues in evaluating UBI, or as if understanding those differences could be straightforwardly extrapolated into an understanding of the probable effects of policy introduced on a national scale.

Researchers usually include caveats about those limitations, but a list of caveats falls far short of a discussion of how the information found relates to the most important questions to ask in evaluating the potential for national adoption of a UBI program.

The potential for the streetlight effect plays a large role when this book considers which questions in the UBI discussion experiments can and cannot address.

5 The Difficulty of Separating the Effects of the Size from the Effects of the Type of Policy Being Studied

If implemented as most supporters envision, UBI involves both a large change in social welfare *strategy* and a large increase in social welfare *spending*. If we want an experiment to help us understand how UBI differs from other strategies, we need to separate the effects of the size from the effects of the type of program being studied.

Haven, CT: Yale University Press, 2014); Angus Deaton, "Instruments, Randomization, and Learning About Development," ibid.

Separating the effects of size and type is extremely difficult in a UBI experiment. The experiments in the United States in the 1970s tested various sizes of NIT, but they only had one control group, all the members of which were eligible for the welfare system existing at the time (see Chap. 6). Thus, the effects of the larger NITs were compared to the effects of the existing system and to smaller NITs, but not to equally generous versions of the existing system. This method gave some information about how the effects of NIT differ by size and some idea about how the effects of NIT differed from the effects of the existing system, but it could not determine the extent to which the effects of the larger NITs had more to do with their being larger or more to do with their being NITs rather than just a more generous version of the existing system.

Furthermore, most reports of results (including those summarized in Chap. 6) lumped together the findings from various experimental groups with various grant levels and marginal tax rates. This amalgamation not only made it difficult to separate the effects of size and type, but also made it difficult to interpret just what size of UBI was being tested on average. What then do the numbers say about the choice between introducing a generous UBI or using the same amount of money to make the existing system more generous or to introduce some other strategy? Unfortunately, it is difficult to extrapolate an answer from the experimental evidence. And that question is far closer to what people most want to know than whether the control group behaves differently from the experimental group. There are two ways to get the estimates closer to what we really want to know.

The first option is to include several different control groups facing differently generous versions of the existing system or whatever system UBI is being tested against. This might seem easy, but to get a really good estimate of the different effects of size and type of spending, each version of UBI would have to be paired with a different strategy of exactly the same size.

Unfortunately, for two so different strategies, it's difficult to determine in advance what size is the same. The cost of a public policy depends on overhead costs, take-up rate, and other factors, most of which can't be estimated in an experiment. Researchers can use data from other sources to estimate what an equal-sized version of the existing system might be. Although any estimate will be highly approximate, just having various sizes for the control groups will help tease out the difference between size and type.

However, none of the NIT or UBI experiments conducted so far have used this technique, and I don't expect any of the currently-under-discussion experiments will either, for one simple reason. It's expensive. It roughly doubles the cost of the experiment. Researchers will have to give out twice as many checks each week, and they will have to deal with the difficult administrative challenge of determining how much each individual in the control group would be eligible for this week if programs A, B, C, and D were X% more generous. They will have to somehow make up the difference, which is probably difficult enough for cash benefits, and extremely difficult for in-kind benefits such as public housing or food stamps.

The second option for examining the difference between size and type is to use theory and data from elsewhere in computer simulations to estimate how the control group would have responded to a more generous version of the existing system and use that as the baseline for comparison or at least as a way to estimate what portions of each observed difference between the control and experimental group are attributable to size or type. This method would also be highly approximate, but nevertheless, it is a potentially useful check on the simple comparison.

I don't know of any literature on past experiments that attempted to use this method. It was not emphasized in the discussion of any NIT or UBI experiment completed so far. Instead most of the literature reported the observed differences between the control group and the experimental group, mentioning what the two groups were eligible for, and sometimes with no further explanation at all, leaving it up to readers to understand that the results, therefore, involve some amalgamation of the effects of the size and type of plan being studied. The popular literature at the time shows little or no awareness of this issue.

The two methods of accounting for the difference between size and type are expensive or difficult or not necessarily very accurate or a mix of all three. Simply explaining the issue takes some effort and all it does is leave readers with the possibly disappointing realization that the numbers are less meaningful than they might initially have appeared to be.