

Open Access

Kate Baldwin and Rikhil R. Bhavnani*

Ancillary Studies of Experiments: Opportunities and Challenges

Abstract: “Ancillary studies of experiments” are a technique whereby researchers use an experiment conducted by others to recover causal estimates of a randomized intervention on new outcomes. The method requires pairing randomized treatments the researchers did not oversee with data on outcomes that were not the focus of the original experiment. Since ancillary studies rely on interventions that have already been undertaken, oftentimes by governments, they can provide a low-cost method with which to identify effects on a wide variety of outcomes. We define this technique, identify the small but growing universe of papers that employ ancillary studies of experiments in political science and economics, and assess the benefits and limitations of the method.

Keywords: ancillary studies; causal inference; downstream experiments; experimental methods; government performance; research design.

JEL Classification: C9; O43; D7.

DOI 10.1515/jgd-2014-0010

1 Introduction

Randomized interventions have substantially advanced our social scientific understanding of the world.¹ But these interventions, and particularly randomized control trials (RCTs, or “field experiments,” as they are called in political science),

¹ For reviews, see Banerjee and Duflo (2009) and Humphreys and Weinstein (2009).

***Corresponding author: Rikhil R. Bhavnani**, Assistant Professor of Political Science at the University of Wisconsin–Madison, 110 North Hall, 1050 Bascom Mall Madison, WI 53705, USA, e-mail: bhavnani@wisc.edu

Kate Baldwin: Assistant Professor of Political Science at Yale University, New Haven, CT, USA

also have drawbacks. They frequently take multiple years to implement and can involve million dollar budgets, causing some scholars to question whether they are worth the cost (Heckman and Smith 1995). Ethical concerns and logistical difficulties also prevent these experiments from addressing some policy questions (Deaton 2010), oftentimes those related to government performance.

However, as experimentation becomes more common in the social sciences and policy evaluation, opportunities are arising for social scientists to use previous experiments to study new outcomes. Researchers can collect data on those assigned to treatment and control groups in previously executed experiments, and then rely on the initial randomization to identify new effects. We refer to this technique as “ancillary studies of experiments.” Ancillary studies can be thought of as using “found” rather than “designed” experiments.

Ancillary studies of experiments provide many of the advantages of randomized interventions, but at lower cost, since the intervention has already been undertaken. In addition, ancillary studies can complement randomized interventions by addressing questions that are difficult for researcher-designed experiments to study.² This is partly because many “found experiments” are not researcher-run RCTs at all. Rather, they are oftentimes lotteries implemented by governments which are less ethically and resource-constrained than individual scholars. As a result, ancillary studies have examined the effects of expensive interventions on sensitive outcomes, despite the fact it would be ethically difficult and logistically challenging for researchers to implement their own interventions to analyse these effects.

Yet along with the great potential of ancillary studies of experiments, this research method has some unique challenges. In this paper, we define and provide an overview of ancillary studies in economics and political science, and analyse the benefits and limitations of this relatively new research method. We begin by defining ancillary studies of experiments. Next, we take stock of the body of research which uses this technique, drawing on a new database of ancillary studies that we make publicly available.³ We then provide a discussion of the logistical challenges of conducting this type of research. We conclude by

² Of course, the converse is also true. As we discuss later, ancillary studies typically examine outcomes after the lapse of some time from the original experiment, so they are best positioned to examine longer term outcomes.

³ The database provides a comprehensive listing of ancillary studies of experiments in economics and political science, along with the characteristics of such studies, including the nature of, and reasons for; the original intervention, the dependent variable, the precise technique used, and a number of other fields. The database allows us to highlight what this research has accomplished, and also its limitations to date. It is available at katebaldwin.commonsworld.org and www.rikhilbhavnani.com.

discussing the potential for increased collaboration between scholars to allow the same randomized intervention to be used to study multiple outcomes.

2 Defining Ancillary Studies of Experiments

Ancillary studies of experiments leverage completed randomized interventions to identify new effects. Once a randomized intervention occurs, it becomes part of the history of the individuals or communities involved. As a result, future scholars can identify new effects by looking for differences across the samples randomly assigned to the treatment and control groups in the initial intervention. The defining characteristics of ancillary studies are that they use a randomized intervention – that the researchers themselves did not usually design or oversee – to study an outcome that was not the primary focus of the original study. As a result, ancillary studies can be thought of as using found rather than designed experiments. Such analyses typically have a time lag between the intervention and the new analysis, and usually involve compiling new data.

While there are many ways of classifying ancillary studies of experiments, in this paper, we distinguish between studies based on government lotteries that randomize a cost or benefit, which are oftentimes conducted for reasons other than evaluation, and randomized control trials, which are run by researchers in collaboration with implementing partners for reasons of evaluation.⁴ Our discussion throughout the paper distinguishes, as necessary, between these two sources of ancillary studies, since the challenges and promise of ancillary studies is frequently conditional on the source of the original randomization.

One of the first examples of an ancillary study of an experiment was conducted by Angrist, who took advantage of the Vietnam draft lottery – a government lottery – to study the effects of military service on lifetime earnings (Angrist 1990). The Vietnam draft lottery has subsequently been used by other scholars to study the effects of military service on everything from economic outcomes and health to criminal behavior and political opinions.⁵

⁴ Other ways of classifying ancillary studies of experiments include whether the intervention was government or researcher-led. When governments and researchers collaborate to evaluate a programme, however, this distinction is hard to make. It does, however, substantially overlap with the distinction between government lotteries and RCTs that we use.

⁵ Angrist and Chen (2008); Angrist et al. (2010); Bergan (2009); Conley and Heerwig (2009); de Walque (2007); Dobkin and Shabani (2009); Eisengberg and Rowe (2009); Erikson and Stoker (2011); Frank (2007); Goldberg et al. (1991); Hearst et al. (1986); Henderson (2010); Lindo and Stoecker (2012); Rohlfs (2010).

Ancillary studies have also been used by scholars to study economic and political outcomes in developing countries. A number of scholars have used the randomized process by which Indian governments have reserved or set aside seats in local legislatures for women to identify the effects of reservations on the chances of women being elected (Beaman et al. 2009, 2012; Bhavnani 2009) and government spending (Chattopadhyay and Duflo 2004).

The studies cited above take advantage of randomizations conducted by governments as means of allocating a cost or benefit.⁶ However, a whole new generation of ancillary studies has been made possible by the increased prevalence of randomized control trials (RCTs) in development economics. These scholar-led trials do not simply allow the initial researchers to identify programme effects. They also open the opportunity for other scholars to assess the effects of the interventions on new outcomes. For example, a number of scholars have used a deworming intervention designed and studied by Miguel and Kremer (2004) to study the long-term effects of deworming (Ozier 2014; Baird et al. 2011). De La O (2013) used the randomized roll out of Mexico's PROGRESA programme to examine the effect of social spending on support for the incumbent, and one of the authors (Baldwin) is part of a team using a randomized evaluation of an NGO's activities in Ghana to estimate the impact of service provision by NGOs on electoral support for incumbent politicians.

All of these studies are ancillary studies of experiments in so far as the researchers leveraged pre-existing randomized interventions designed and overseen by other scholars or policy makers to study outcomes not considered by the original studies. The first wave of research followed up on government-run experiments, while a newer wave is building on scholar-led RCTs. As we discuss below, both sets of studies share the challenge of leveraging the interventions to answer new questions, although the challenges and benefits sometimes apply in different degrees to government-run lotteries and RCTs.

We introduce the term “ancillary studies of experiments” because it describes a hitherto unrecognized subset of experimental analyses that share an approach and face a common set of challenges.⁷ Ancillary studies include a subset of “natural experiments,” defined as data that come from naturally occurring phenomenon that are not under the control of the analyst but in which assignment to the treatment and control is random or “as-if” random (DiNardo 2008; Dunning

6 In the case of the Vietnam draft lottery, this is the cost of serving in the military. In the case of the reservation lotteries, this is the cost to male incumbents of not being able to re-run for office.

7 We thank Don Green and the journal editors for helping to develop this terminology.

2012; Sekhon and Titiunik 2012).⁸ We exclude studies that rely on as-if randomization from our definition (such as experiments that rely on as-if randomization due to nature – see Rosenzweig and Wolpin 2000) to focus on found experiments that are explicitly randomized. Ancillary studies of experiments are a broader set of studies than “downstream experiments” as originally conceived by Green and Gerber (2002). As originally defined, downstream experiments use historical randomized interventions as *an instrument* to identify the effect of the original outcome on another variable of interest (Green and Gerber 2002: 394). In contrast, ancillary studies include both downstream analyses and analyses that consider the direct effect of the original treatment on new outcomes.⁹

3 Taking Stock

In this section, we take stock of the use of ancillary studies of experiments to date in economics and political science, drawing on a new database of ancillary studies. The database includes both published research and working papers. It was constructed in three steps. First, we searched social science databases using key word searches.¹⁰ Then we emailed organizations and listservs in the relevant subfields of economics and political science. Finally, we used snowball sampling, using the citations of and in the identified ancillary studies to search for additional studies. Because we found that ancillary studies often clustered around large randomized interventions, we also searched for articles that mentioned

8 This is the definition of natural experiments that is currently widely accepted in political science and economics. Harrison and List’s earlier definition of natural experiments (2004) also restricted focus to truly randomized treatments, but did not encompass analyses building on experiments conducted by other researchers.

9 There are some analytic issues specific to analysing downstream experiments related to instrumental variable estimation which we do not discuss in this essay. For a review, see Sondheimer (2011).

10 The databases consulted were the Social Science Research Network (SSRN), the Social Sciences Citation Index (SSCI), Social Sciences full text, Web of Science, JOLIS, JSTOR, Cambridge journals online, British Library for Development Studies, IDEAS Economic and Finance Research, ScienceDirect, Sage full-text collections, C2 SPECTR, Google Scholar, and Google. Searches were conducted between July and October 2012. The key words we searched were: “Downstream Experiment” or [“Natural Experiment” and (“Random” or “Randomized” or “Randomization” or “Randomised” or “Randomisation” or “Lottery” or “By lot” or “Drew lots”)] or [(“Completed or Old or “Previously conducted”) Field Experiment”]. Searches were conducted in English. Our database is comprehensive if all ancillary studies of experiments were included in the databases we searched and if they all used at least one of our search terms.

each of the randomized interventions used in the identified ancillary studies. Full details on the protocol for creating the database and the database itself are available at the sites mentioned in footnote 3.

Studies were coded as being ancillary studies if they used a randomized intervention that was not designed by the researchers to measure effects on a new outcome. As a result, we did not count replication studies of the original experiment as ancillary studies.¹¹ We also did not include studies that used natural experiments due to geography, regression-discontinuity designs or other as-if randomized interventions. To qualify as an ancillary study of an experiment, the study needed to assign units to treatment and control groups via an intentionally random process, as is the case in RCTs and government lotteries.¹²

Ancillary studies are distinguished from government lotteries or RCTs in that the latter are explicitly designed to examine the outcomes of interest while the former use experiments to study effects beyond those intended in the initial design. When it was difficult to determine whether or not the scholars designed the experiment or not, we coded these papers as ancillary studies. It was sometimes also difficult to determine ex post which outcome variables an RCT or government lottery was initially intended to analyse. We classify studies as ancillary studies only if they examined an outcome not included in the reports written by the initial research team.¹³

As a result of our research, we found 82 studies that qualify as ancillary studies of experiments. Table 1 lists these, organizing them by the substantive area of treatment.

It is noteworthy that 43 of 82 ancillary studies in the database relate to government performance, if we define these studies as those whose dependent

11 We do not consider papers to be ancillary studies of experiments if they simply check the robustness of the initial analysis to different specifications or sub-group analyses. They must look at different outcomes or the same outcomes over different time periods or in different locations. Our database will necessarily have some incorrect inclusions and exclusions, since we can rarely ascertain exactly what researchers initially intended to do.

12 Some famous lottery studies do not actually meet these criteria. For example, the Imbens et al. (2001) study of the effect of unearned income from lotteries compared lottery winners who won different amounts across different lotteries. Because the group of people who play each lottery differs, assignment to different levels of the treatment is not necessarily random. Of course, even RCTs that assign populations to treatment and control groups via truly random processes sometimes fail to achieve balance on certain variables or experience imperfect compliance. This did not disqualify a study from being included in the database. We also include at least one study that claimed to use randomization but whose assignment method was subsequently criticized as not truly random. See the discussion of Kremer and Miguel (2004) in Deaton (2010).

13 As a result, we may be excluding some ancillary studies in which the initial research team developed new uses for their experiment ex post.

Table 1: Summary of Ancillary Studies of Experiments Database.

Randomized intervention	Study citations	Number of studies
Access to funds/loans	Agarwal et al. 2010 Bagues and Esteve-Volart 2011 De La O 2013 Hite 2012	4 studies
Co-worker characteristics	Guryan, Kroft and Notowidigdo 2009 Rogowski and Sinclair 2012	2 studies
Educational services	Angrist et al. 2002 Chetty et al. 2011 Cullen et al. 2006 Dynarski et al. 2011 Hastings et al. 2007 Hemelt et al. 2013 Krueger and Whitmore 2001 Rouse 1998 Sondheimer and Green 2010	9 studies
Evaluation committee characteristics	Bagues and Perez-Villadoniga 2012 De Paola and Scoppa 2011 Zinovyeva and Bagues 2011	3 studies
Health services	Doyle et al. 2010 Baird 2007 Baird et al. 2011 Hoddinott et al. 2008 Hoddinott et al. 2013 Li et al. 2003 Maluccio et al. 2009 Ozier 2010 Pollitt et al. 1995 Stein et al. 2008	10 studies
Housing services	Gay 2012	1 study
Immigration/Visas	Clingingsmith et al. 2009 Gibson et al. 2009, 2010, 2011 Gibson et al. 2010 McKenzie et al. 2006, 2007a, 2007b Stillman et al. 2006 Stillman et al. 2012	10 studies
Judge characteristics	Abrams and Yoon 2007 Green and Winik 2010 Kling 2006 Sen 2012	4 studies

Randomized intervention	Study citations	Number of studies
Military service	Angrist 1990 Angrist and Krueger 1992 Angrist and Chen 2008 Angrist et al. 2010 Bergan 2009 Conley and Heerwig 2009 de Walque 2007 Dobkin and Shabani 2009 Eisenberg and Rowe 2009 Erikson and Stoker 2011 Frank 2007 Gallani et al. 2011 Goldberg et al. 1991 Hearst et al. 1986 Henderson 2010 Lindo and Stoecker 2012 Rohlf s 2010 Siminski and Ville 2012	18 studies
Political information	Ferraz and Finan 2008	1 study
Political power/position	Brockman and Butler 2012 Gaines et al. 2012 Ho and Imai 2008 Kellerman and Shepsle 2009 Loewen et al. 2014	5 studies
Reservation of political seats	Beaman et al. 2009 Beaman et al. 2012 Bhavnani 2009 Chattopadhyay and Duflo 2004	4 studies
Roommate characteristics	Barnhardt 2009 Boisjoly et al. 2006 Duncan et al. 2005 Han and Li 2009 Foster 2006 Kremer and Levy 2008 Sacerdote 2001 Stinebrickner and Stinebrickner 2006, 2007 Van Laar et al. 2005 Yakusheva et al. 2011	11 studies

Source: Ancillary studies of experiments database compiled by authors. See text for details.

variables have to do with state provided goods and services, including education, health and justice, and outcomes that the state takes responsibility for, such as income. This count is even higher – at 59 studies – if we code all studies based on government-led interventions as pertaining to governmental performance. While scholars have complained that RCTs do not easily lend themselves to the evaluation of governance-related interventions (Rodrik 2009; Deaton 2010), ancillary studies appear better able to do this.

As suggested previously, one reason why ancillary studies of experiments are so prevalent in the field of governance is that they build on randomized interventions conducted for two different purposes. In some cases, they are structured around completed RCTs conducted for reasons of evaluation. For example, this was the reason why aspects of Mexico's PROGRESA programme were randomized.¹⁴ However, they also build on lotteries conducted by governments for reasons of fairness. In cases where it is not possible to distribute a benefit (or a cost) to all, randomization avoids discrimination by giving everyone the same chance of being chosen. This was the rationale for drafting men to the United States (US) military by lot during both the First and the Second World Wars, and the Vietnam War,¹⁵ and also purportedly for randomly reserving electoral seats for female candidates in India. The vast majority of the ancillary studies we identified rely on lotteries conducted for reasons of fairness which partly explains the prevalence of ancillary studies in the study of governance. However, it also suggests that RCTs have been largely untapped as a source of ancillary studies, a fact to which we will return in our critical assessment of the field's accomplishments.

Although ancillary studies of experiments have had some success in studying phenomena – such as large government interventions – that are not easily amenable to RCTs, the method also has limitations in the substantive areas to which it has been applied to date. As Table 1 makes clear, to date, most ancillary studies have been built on just a few *types* of interventions. Even more specifically, there has been a large amount of clustering around specific interventions.¹⁶ For

14 Bureaucrats hoped that definitively demonstrating the efficacy of the programme would reduce the risk of the programme being eliminated with changes in government (Parker and Teruel 2005: p. 208).

15 President Johnson stated in a special message to the Congress prior to the establishment of the Vietnam draft, “The paramount problem remains to determine who shall be selected for induction out of the many who are available... I have concluded that the only method which approaches complete fairness is to establish a Fair and Impartial Random (FAIR) system of selection which will determine the order of call for all equally eligible men.” Quoted in Fienberg (1971: p. 255).

16 This clustering was obvious even before the third part of our search protocol, which searched for studies mentioning the randomized interventions that had been used in previously identified ancillary studies.

example, 22% of the studies are based on draft lotteries, of which almost 90% use the Vietnam draft lottery. Another 12% of studies are based on international visa/immigration lotteries, of which 90% use the Tonga–New Zealand immigration lottery. Multiple studies have also used the government of India’s randomized reservation of seats for women, the STAR classroom-size experiment in Tennessee, the INCAP nutritional supplement experiment in Guatemala, and the Kremer and Miguel deworming experiment in Kenyan schools to examine new outcomes. This raises concerns both about the breadth of applicability of the method and the external validity of the findings of these studies.

We return to these concerns in the next section where we suggest directions for future research that would partly alleviate these limitations in how ancillary studies have been applied to date. In considering the strengths and weaknesses of this body of research, it is important to recognize that ancillary studies of experiments are a very recent phenomenon. The first study in our database is from 1986, and more than half of all studies have been produced in the last 5 years.¹⁷ Ancillary studies have just begun to be explored as a research method, and much more can be done with this research technique.

3.1 What has been Accomplished to Date

The accomplishments of ancillary studies of experiments to date fall into two main categories. First, they have demonstrated themselves to be a relatively low-cost technique for identifying empirical effects. Second, they have proved able to examine effects that RCTs have had difficulty studying for logistical and ethical reasons.

Since ancillary studies of experiments do not incur any of the costs involved in designing an experiment, they are a relatively low-cost research technique. The ancillary studies database demonstrates this. Although ancillary studies can involve a wide variety of different data collection techniques, with a wide range of associated costs, most (51 of 82) of the papers in our database collected data on

¹⁷ We tried to be as expansive as possible in the search terms that we used. Because the trends we measure parallel the expansion of experiments more broadly in economics and political science, we doubt that these findings are an artifact of our search terms. For example, Gerber (2011) notes that “field experiments” (experiments conducted in natural settings in which participants are not aware of the experiment) were rare in political science before 2000, with no political science journal publishing research based on a field experiment in the 1990s. See also the discussion of trends in development economics in Banerjee and Duflo (2009).

new outcomes from government records or “off-the-shelf” surveys. Only 31 of 82 studies involved expensive follow-up surveys designed by ancillary researchers.

Relatedly, the database shows that it is possible for the same intervention to be used to study a wide variety of outcomes within and across disciplines. The Vietnam draft has been used to study the impact of serving in the military (or expecting to serve in the military) on economic outcomes,¹⁸ health outcomes,¹⁹ violence and criminality,²⁰ and political attitudes.²¹ The Tonga–New Zealand migration lottery has been used to study the impact of migration on the income of the migrating family members,²² the income of those left behind,²³ and the physical and mental health of the migrants.²⁴ Various roommate studies have analyzed the impact of peer effects on inter-racial or inter-religious attitudes,²⁵ drug and alcohol use,²⁶ educational outcomes,²⁷ and weight gain.²⁸ This indicates possibilities for cost reduction and cost-sharing among scholars interested in a wide variety of substantive outcomes. The database also includes examples of scholars collaborating across disciplines to study the effects of a particular intervention, an exciting development insofar as it is likely to allow the exchange of knowledge and best practices across disciplines.²⁹ Ancillary studies are plausibly particularly amenable to cross-disciplinary research since the costs of the original intervention have already been borne, and since the search for additional outcomes can lead people outside their disciplinary homes.

18 Angrist (1990); Angrist and Krueger (1992); Angrist and Chen (2008); Frank (2007).

19 Angrist et al. (2010); Conley and Heerwig (2009); de Walque (2007); Dobkin and Shabani (2009); Eisenberg and Rowe (2009); Goldberg et al. (1991); Hearst et al. (1986).

20 Lindo and Stoecker (2012); Rohlfs (2010).

21 Bergan (2009); Henderson (2010).

22 McKenzie et al. (2006).

23 Gibson et al. (2009, 2010a); McKenzie et al. (2007).

24 Gibson et al. (2010b, 2011); Stillman et al. (2006, 2012).

25 Barnhardt (2009); Boisjoly et al. (2006); Van Laar et al. (2005).

26 Duncan et al. (2005); Kremer and Levy (2008).

27 Foster (2006); Han and Li (2009); Sacerdote (2001); Stinebrickner and Stinebrickner (2006); Stinebrickner and Stinebrickner (2007).

28 Yakusheva et al. (2011).

29 For example, economists and biomedical researchers collaborated on the 40-year follow-up on the INCAP nutritional experiment in Guatemala, and they explicitly emphasized the benefits of a multidisciplinary follow-up team in their grant application to the National Institute of Health (NIH). See the discussion in Martorell et al. (2005) for more details. For other examples of interdisciplinary research teams, see Boisjoly et al. (2006) and Duncan et al. (2005). In another example, Hite (2012), a political scientist, used ethnographic methods to follow-up on a study run by economists Karlan and Zinman (2009).

The second achievement of ancillary studies of experiments has been their ability to study large-scale interventions and sensitive topics. Many (59 of 82) of the original experiments in the database have been implemented by governments. As a result, ancillary studies provide a useful complement to government interventions and RCTs, the vast majority of which rely on interventions implemented by NGOs (Bruhn and McKenzie 2009). The conclusions from the RCT revolution in development economics have been criticized on the grounds that the results from evaluations implemented by small, carefully selected NGOs may not apply to interventions conducted on a larger scale by governments, due to general equilibrium effects, lower capacity, or greater corruption (Barrett and Carter 2010; Deaton 2010). In view of these concerns, the fact that more than 70% of ancillary studies examine government-implemented interventions is an advantage. Ancillary studies can provide important tests of how well programmes scale and are executed by the public sector.

Relatedly, ancillary studies of experiments often permit the systematic study of interventions that ethics would not allow to be randomized for reasons of evaluation, but that governments have decided should be randomized for reasons of fairness. For example, it would not be considered ethical for researchers to design an experiment randomizing military service or incarceration. However, governments have run lotteries that effectively do this by randomly pulling draft numbers and randomly assigning defendants to lenient and harsh judges, and scholars have used these government-run lotteries to measure the effects of serving in the military (Angrist 1990) and being incarcerated (Kling 2006).

Finally, even when those who conduct ancillary studies build on RCTs, they are often able to study topics that the initial researchers could not. This is because the ethical burden of observing the outcomes that follow from an intervention are different from the ethical burden of manipulating an intervention for the purpose of creating a particular outcome. For example, it may be considered ethically problematic to manipulate conditional cash transfers with the express purpose of studying whether they affect support for a particular political party. However, if access to conditional cash transfers has been randomized for other reasons (such as studying poverty alleviation), there may be fewer concerns about conducting a follow-up study on the intervention's political effects. In addition, scholars can use an instrumental variables framework to estimate the effects of variables that it would not be ethical to randomize. For example, SONDHEIMER and GREEN (2010) use exposure to educational programming as an instrument for the effect of education on voter turnout.

Of course, building on government-led programs poses its own set of ethical dilemmas, and the fact that a randomized program is government-run does not give researchers *carte blanche*. In particular, scholars that work with “found”

rather than designed experiments should consider both the ethics of their own data collection methods and the consequences of their study for the original intervention.³⁰ Still, there will often be room for ancillary studies to study sensitive topics while meeting high ethical bars.

The fact that ancillary studies of experiments rely on found experiments that they do not bear the cost of designing has made them particularly useful in the study of governance. They have been able to study large-scale government interventions, such as draft lotteries or the implementation of reserved seats for women. In addition, they have been able to study politically sensitive topics, such as the effect of preferred access to government services on levels of incumbent political support and political participation (Hastings et al. 2007; De La O 2013). Because RCTs have often found it difficult to study these types of phenomena, these are particularly important accomplishments.

3.2 What Remains to be Accomplished

Although ancillary studies of experiments allow researchers to examine more sensitive and large-scale effects at lower cost than is typically the case with government lotteries and RCTs, ancillary studies are by no means a panacea to the shortcomings of experimental methods. Governments may face more relaxed resource and ethical constraints than academics but they are not unconstrained. The clustering of ancillary studies around particular interventions and issues is indicative of such constraints. In this section, we briefly discuss some shortcomings of the corpus of ancillary studies documented previously.

A striking pattern in our review of ancillary studies of experiments is the scant number of studies that replicate the findings of other ancillary studies. By our count, five of the 82 ancillary studies in the database were replications. There is a need for greater replication of ancillary analyses of particular effects in different settings. In many ways, it is surprising that there has not been more of this to date, as the database suggests strong demonstration effects in the search for randomized interventions: once one scholar has identified an intervention that was randomized in one instance – for example, military drafts, roommate assignments, positions on academic promotion committees, or judge assignments – other scholars find other examples of similar interventions being randomized. However, for the most part,

³⁰ For example, if a program was initially conceived as non-partisan, there may be ethical considerations if questions are framed to make individuals believe it was partisan. For a broader discussion of the ethics of RCTs more generally, see Humphreys in this volume.

scholars have used different examples of the same type of intervention to study different effects, rather than trying to replicate the effects from the first study.³¹ Future research should prioritize the replication of ancillary studies in different settings, through stand-alone follow-up studies or by incorporating results from multiple settings in the initial publication. Publications based on ancillary studies would appear particularly well-suited to incorporate replications across multiple sites because this research method requires less investment of time and resources compared to researcher-designed RCTs. For example, it would be possible for the same scholar to examine the health effects of military drafts in the US, Argentina, and Australia. In one promising example, Hite-Rubin is currently in the process of replicating an earlier ancillary study that she conducted on the effects of access to credit on political orientations in the Philippines with a group of researchers who conducted a similar credit experiment in Mexico.³²

Relatedly, this area of research appears to have many randomization-driven searches for questions, but few theory-driven searches for randomizations. Of course, it is difficult to determine whether the question or the data motivated the research project. However, there are many examples of the same set of authors using one intervention to study multiple outcomes which strongly suggests a data-driven process. The most obvious example is the set of papers written by Gibson, McKenzie, and Stillman using the Tonga–New Zealand lottery to study everything from economic outcomes to mental health. In contrast, if research is driven by theoretical questions, we would expect more papers that use multiple examples of the same type of intervention to measure the effects of this intervention on one outcome. There is only one example of this in the data set, the article by Sondheimer and Green (2010) on the effects of education on voter turnout. In this case, it is obvious that the authors started with the question and then searched for all available studies that would allow them to answer this question. More future studies should follow this best practice.

Finally, surprisingly few (16 of 82) ancillary studies have built on RCTs. Instead, most (66 of 82) studies build on interventions that were randomized by

31 A partial exception has been the replication of studies that examine the effect of peer academic achievement on students' own grade point averages (GPAs) using roommate randomization. These have been replicated across several universities and at least three different countries (Sacerdote 2001; Foster 2006; Stinebrickner and Stinebrickner 2006, 2007; Han and Li 2009). But even in this case, more of the studies inspired by the original Sacerdote (2001) study have used other examples of roommate lotteries to study new outcomes, such as inter-racial attitudes (Van Laar et al. 2005; Boisjoly et al. 2006), drug use and sex (Duncan et al. 2005), alcohol use (Kremer and Levy 2008), and weight gain (Yakusheva et al. 2011).

32 Angelucci et al. (2015).

governments or others for reasons of fairness. This has provided a useful counterpoint to RCTs which have been limited in their study of government-run interventions. However, it has probably contributed to the restricted substantive scope of ancillary studies to date, the limited replication of ancillary studies, and the rarity of question-driven searches for randomizations because an enormous source of randomized interventions has been mainly unexploited. Notable exceptions are the group of studies examining the long-term effects of the STAR experiment (Krueger and Whitmore 2001; Chetty et al. 2011; Dynarski et al. 2013), the group of studies examining the educational and economic impact of the INCAP nutritional experiment (Pollitt et al. 1995; Li et al. 2003; Hoddinott et al. 2008, 2013; Stein et al. 2008; Maluccio et al. 2009), and a set of three studies that build on the initial Kremer-Miguel deworming study (Baird 2007; Ozier 2014; Baird et al. 2011).³³ Similarly, Hite (2012) piggybacked on a microfinance experiment run by Karlan and Zinman, and one of the authors (Baldwin) is currently conducting research based around an evaluation of an NGO's service provision activities run by Karlan and Udry. De La O (2013), Gay (2012), and Sondheimer and Green (2010) build on bigger evaluations of government programmes. However, when one considers the sheer magnitude of the number of randomized control trials that have been run in development economics during the past decade (the American Economic Association's RCT registry lists 287 RCTs in 64 countries), it is surprising that there have not been more ancillary uses of these interventions. The possibility for collaboration across different sub-fields and even different disciplines in this area is great but largely untapped.

4 How to Create an Ancillary Study of an Experiment: Major Challenges

While ancillary studies of experiments are a new and exciting frontier for research, they are subject to a number of challenges. Some of the challenges of ancillary studies are shared by experimental designs in general (including compliance and spillover problems), and are well-covered elsewhere.³⁴ Other challenges are shared with natural experiments, although ancillary studies avoid the

³³ There have been many more re-analyses of the STAR intervention, some of which have focused on the program's effects on particular subgroups. Unless a study examined effects on an outcome variable not considered in the reports by the initial research team, it is not considered an ancillary analysis for our purposes.

³⁴ Duflo et al. (2007), Gerber and Green (2012).

largest difficulty for this research method by excluding studies based on as-if random interventions. We focus on four challenges that are particularly relevant when conducting ancillary studies based on found randomized interventions: these are the matching of social scientific questions to randomizations, collecting information on the randomization scheme, measuring outcomes, and mechanism testing.

4.1 Matching Social Scientific Questions to Randomizations

The first challenge for a scholar interested in crafting an ancillary study is finding a pre-existing randomized lottery that speaks to a social scientific question of interest. Unlike scholars designing their own randomized experiments, who generally develop their design to answer specific questions, researchers hoping to conduct an ancillary study may start with a research question but then find only an imperfect match between a pre-existing experiment and their ability to answer that question, or they may stumble upon a randomized intervention before they have clearly articulated their research question of interest. In either case, a clear question that speaks to theoretical debates needs to be fashioned.³⁵ This is the first order of business, and demands creativity.

Perhaps the easiest place to find a randomized study is the database of ancillary studies of experiments introduced previously.³⁶ The randomized interventions that these studies draw on have all been successfully redeployed to study ancillary outcomes. Scholars may additionally look at the increasing number of government, NGO, and donor-led interventions in which treatments were randomized. The American Economic Association's RCT registry, for example, lists 287 RCTs in 64 countries. Many (59 of 82) ancillary studies of experiments have employed lotteries run by governments, but the RCT revolution in development economics and the increasing number of donors pushing for rigorous evaluations have resulted in a dramatic increase in interventions that are randomized for research purposes. The American Economic Association's RCT registry, the Economics Research Network (ERN) Randomized Social Experiments e-journal and the web sites for the Abdul Latif Jameel Poverty Action Lab (J-PAL), and Innovations for Poverty Action, the leading organizations in the field of randomized

³⁵ A similar point was made by Rosenzweig and Wolpin (2000) in the conclusion of their review of "natural natural experiments" (as-if randomizations caused by nature).

³⁶ As discussed, this risks resulting in even more studies built around a small number of interventions.

evaluations in economics, provide fairly comprehensive listings of on-going and recently completed RCTs. Many of these RCTs offer opportunities for ancillary studies, but they also raise questions about norms of experiment-sharing, an issue to which we return in the final section.

Of course, not all randomized interventions will lend themselves to ancillary studies. Large-scale randomized interventions that have substantial short and long term effects are more likely to yield ancillary studies. Relatively unobtrusive interventions, which have small immediate impacts, are less amenable, as it will be difficult for scholars who find these experiments after-the-fact to be able to measure effects during the relevant period.³⁷ Still, interventions that are found to have small immediate impacts in one domain may have longer-term outcomes in another domain; for example, it is conceivable that receiving a one-time tax break from the government has little impact on long-term income but greater effects on political views.³⁸

Another concern is that developments between the original intervention and the present could “swamp” any effects of the randomized intervention. For this reason, experiments involving randomized roll-outs will not always be suitable for ancillary analysis.³⁹ Care needs to be taken to understand the degree to which actions in the intervening period affect the original randomization. This is likely to be more of a problem as the time lapse between the original intervention and the present grows. Panel attrition poses a well-known threat to randomization but so do new interventions explicitly conditioned on the original intervention. Studies of the effect of randomized military deployment, for example, will have difficulty separating the effects of military deployment from the effects of receiving veteran’s health care, because the two interventions are bundled. One way around this is to reframe the paper as investigating the effect of the bundle of interventions (in this example, military service and veteran’s healthcare), or, even more simply (since we oftentimes do not know the entire contents of the bundle), as the effect of the original lottery itself (the Vietnam draft).

Once a new question has been matched to a randomized intervention, scholars have to ensure that the randomization is valid. Doing so entails investigating

³⁷ When scholars can use historical data to measure their outcome of interest, they may also be able to investigate the short-term impact of the intervention in an ancillary study. It is noteworthy that most (51 of 82) of the papers in our database collected data on new outcomes from government records or “off-the-shelf” surveys.

³⁸ The example is from private correspondence with Thad Dunning.

³⁹ Exceptions are ancillary studies in which the length of exposure to the treatment is theoretically relevant. For examples, see Baird (2007), Baird et al. (2011), and Ozier (2014).

the integrity of the original randomization. Was the lottery carried out properly?⁴⁰ How were exceptions dealt with?⁴¹ And Are the resulting treatment and control groups, in fact, balanced in terms of pre-treatment covariates?⁴² While the original research may have reported balance on the pre-treatment covariates most pertinent to the initial experiment, the switch to a new outcome measure in most ancillary studies will typically suggest new pre-treatment covariates on which to check for balance.

In addition, scholars conducting ancillary studies of experiments need to carefully consider the population over which the randomization occurred, and the implications this has for the scope of their findings. Unlike in experiments that are fully under the control of the experimenter, the scope conditions for ancillary studies are determined by the original intervention, and not the experimenter. Oftentimes, this means that the population that the ancillary studies can speak to is narrower than the scholar would like. An example of this is Bhavnani's (2009) study which examines the effects of the randomized reservation of seats for women in elections in 1997, on the chances of women winning office in the subsequent open elections in 2002. Since reservations for women have been in place in the context studied since 1992, the uncovered effects are contingent both on the existence of a previous round of reservations, and on the concurrent (randomized) use of quotas in other seats in 2002.⁴³

Scholars should also consider the statistical power of the original intervention to identify effects on the new outcome of interest. The effects of the randomized variable on the new outcome may be anticipated to be smaller or larger

40 Lotteries are sometimes not perfectly carried out. Dobkin and Shabani (2009), for example, note that the Vietnam draft lottery in December 1969, was subject to a mechanical failure as the balls were not adequately mixed.

41 The randomizations used by a number of studies in our database had exceptions. In a school voucher lottery study, for example, Angrist et al. (2002) note that in "a few" cities vouchers were sometimes assigned "based on pupils' primary-school performance instead of randomly" (see footnote 5). Abrams and Yoon (2007); Green and Winik (2010); Doyle et al. (2010); and Sen (2012) also note exceptions to the randomized assignment to treatment. Ad hoc assignments pose a threat to the randomized inference, and need to be dealt with, if possible, in detail.

42 While these groups are balanced in expectation, the particular lottery that was conducted need not have resulted in balance.

43 Sekhon and Titiunik (2012) formalize these assumptions which are verbally described in the original paper. Another example of this is Hastings et al. (2007) which examines the effects of school lotteries on voter participation in school board elections. Since lotteries are only used when schools are oversubscribed, the results of this study are only applicable in these places.

than the effects in the initial study, and so the statistical power of the study to identify the relevant effect size is likely to be different.

4.2 Collecting Information on the Randomization Scheme

A second major difficulty for scholars hoping to conduct an ancillary study is to collect details on the randomization. Scholars need to know the probability of each unit receiving the treatment (or simply that the probability was equal for all units) and the treatment each unit was actually assigned.⁴⁴ When there are problems of non-compliance (which might be greater as the time lag between the original intervention and the new outcome being measured increases), details on compliance will also need to be collected.

Experiments in which randomization was done by public lottery will generally be more amenable for ancillary study because it is easier to recover treatment assignment. In addition, it is usually easier to obtain this information in the case of government-run lotteries than it is in the case of researcher-run RCTs because the latter might be constrained by confidentiality agreements. In both instances, accessing the randomization scheme is likely to be particularly difficult when the initial treatment is randomized at the individual rather than the cluster level.⁴⁵

Government lotteries, including those involving public officials, are particularly amenable to ancillary study. For example, Bhavnani's (2009) study could easily recover each unit's treatment probabilities because the lottery was run by the government and every electoral district had an equal probability of being selected to be reserved for women. In the case of the Vietnam draft lottery, ancillary study has been possible because the randomization was run by the government, but was not truly at the individual level. Instead, participants were called by randomly chosen birthdates, information that is more easily obtainable.⁴⁶ In a

⁴⁴ For discussions of how to uncover average treatment effects in the case of unequal treatment assignment, see Duflo et al. (2007), Gerber and Green (2012) and Humphreys (2009).

⁴⁵ As an anonymous reviewer pointed out, this suggests the loss of power that results from public cluster-level lotteries should be weighed against the potentially value of these lotteries beyond the initial study.

⁴⁶ Even in this case, scholars often have to go to some lengths to obtain access to the birthdates of the individuals in their study, because birthdays are often scrubbed from publicly released survey data and records for reasons of confidentiality. See Angrist (1990); Dobkin and Shabani (2009); and Erikson and Stoker (2011).

number of the other studies in our database, the randomization involved a government official (5 of 82) or judge (3 of 82) being assigned a particular power. In both of these instances, there are no confidentiality concerns because of the public status of the units being randomized.

Despite the challenges of recovering individual-level randomizations of non-public figures, many of the ancillary studies of experiments identified in our database do employ such interventions. Sharing data may be easier if the scholar conducting the ancillary study contacts the individuals responsible for the original study before it is complete. An interesting example is Hite (2012), who piggy-backed on a credit-access RCT to examine how access to formal finance impacts the political views and activities of small-business owners.⁴⁷ Field work for the study involved face-to-face interviews with over 200 of the original experimental participants. In order to conduct this research, IRB approval was required, both to access the data from the original experiment, and for follow-up ethnographic field work that involved locating and recruiting original respondents for face-to-face interviews.⁴⁸

Furthermore, scholars are sometimes able to recover information on the assignment of private individuals to different treatments from the government or organizations that ran the lottery. For example, Clingingsmith et al. (2009) were provided data on the names, addresses, and telephone numbers of all the applicants to the 2006 Hajj lottery by the Pakistani government. In other cases, scholars have been provided information on individual-level treatment assignment only after agreeing to conditions designed to protect respondent confidentiality. For example, Sondheimer and Green (2010) were given information on the names and treatment assignment of participants in two educational experiments in the US after signing agreements not to contact the participants and to keep the participants' information confidential.⁴⁹ They were then able to match participants' names to public voting records. In situations where information on the outcome variable is available for the entire population from which the original sample was drawn, another solution is to have the original investigator merge the data file containing the new outcome with the data file containing participants' names and assignment information.⁵⁰ Confidentiality concerns

⁴⁷ For the results of the original RCT, see Karlan and Zinman (2009).

⁴⁸ Private correspondence with an anonymous reviewer.

⁴⁹ Private communication with Donald Green. See also Kremer and Levy (2008).

⁵⁰ Sondheimer and Green (2010) used a similar strategy – hiring a third party to merge the data – in their quasi-experimental analysis of a third educational programme. See also Gay (2012).

make ancillary studies of individual-level randomizations more challenging but not impossible.

Finally, ancillary studies of experiments face the challenge of collecting information on compliance with treatment assignment. Information on treatment assignment is sufficient to calculate the intent-to-treat (ITT) estimate, but in instances with high levels of non-compliance, this may not provide a meaningful estimate of the effects of the intervention. A number of ancillary studies, including the Vietnam draft lottery studies, have not been able to collect information on treatment take-up, but have still been able to generate estimates of the complier average causal effect (CACE) by using other data sources to estimate the proportion of “always takers” and the treated who take up treatment.⁵¹ Alternatively, Erikson and Stoker (2011) managed to turn this problem into an advantage by framing their study as the effects of expected military service on political attitudes.

4.3 Measuring Outcomes and Estimating Effects

Another challenge is to measure the outcome(s) of interest in the ancillary study. Given the time lag between the original experiment and the ancillary study, this often takes significant legwork. For example, in order to examine the impact of educational experiments from the 1960s and 1980s on voter turnout in 2000, 2002, and 2004, Sondheimer and Green (2010) did “years of detective work tracking down the subjects in these studies” (Sondheimer and Green 2010: 176). Such exercises also require the continued consent of subjects for the study of new outcomes.⁵²

Furthermore, in some (9 of 82) ancillary studies, the outcome in which the scholar conducting the ancillary study is interested is measured in a different unit than the unit of randomization. For example, in De La O’s study of the electoral impact of PROGRESA, the randomization was conducted at the village level, but her outcome of interest – support for the incumbent – was available only at the polling precinct level. One of the authors (Baldwin) has faced similar difficulties in analysing the effects of NGO activities on electoral results in Ghana.

The difficulties here are greater than the difficulty of figuring out how the units at which randomization occurred and those at which ancillary outcomes

⁵¹ See Angrist (1990) for an example.

⁵² To this end, the original consent scripts for the study should be examined.

are observed line up with each other, which by itself is often a time-intensive undertaking. The problem is that the new units may have differential probabilities of assignment to treatment than the original units. For example, in De La O's study, all of the villages in the PROGRESA experiment had the same probability of being part of the treatment group. However, the polling precincts – the units at which election results were observed – contained different numbers of villages in the PROGRESA experiment (most contained one village from the PROGRESA study, but some contained two) and different numbers of non-experimental villages (De La O 2013). Thus, the probability of a polling precinct being exposed to different treatment doses differed depending on the number of experimental villages in the precinct. A similar problem emerges if ancillary studies seek to examine second-hand exposure to a technology, such as the effect of a health intervention on the parents or siblings of the children randomly exposed to the intervention. In this case, the probability of assignment to the treatment is correlated with the number of children or siblings in the experimental group.⁵³

At least two solutions to the imperfect overlap problem are possible. One solution is to use surveys to collect data on the ancillary outcomes at the level at which the treatment was randomized. However, this will not always be possible (or perhaps even desirable for some types of data, given recall biases). Survey fatigue might also be an issue here, as the same populations may be surveyed repeatedly if multiple scholars use the same randomization to study different outcomes. An alternative solution is to directly take into account the characteristics of the ancillary units that condition their probability of exposure to the treatment. Researchers can identify the effect of receiving treatment by stratifying ancillary units according to their probability of receiving treatment (De La O and Rubenson 2010). For example, De La O is able to identify the effect of PROGRESA on vote returns by separately analysing precincts with different numbers of experimental villages. In addition, in cases in which units differ between the original experiment and the ancillary study, units in the ancillary study may receive different treatment dosages. In De La O's study, she accounts for different dosages by controlling for the number of villages in each precinct.

The estimation of effects in ancillary studies of experiments also raises some problems of statistical inference and multiple comparisons. Individual studies are increasingly cognizant of the fact that an intervention is likely to be found to have at least one positive effect if enough dependent variables are included in the

⁵³ See the discussion in Humphreys (2009).

study; if scholars “fish” for positive effects by examining the effects of an intervention on 20 different outcomes, they are likely to find one effect that is statistically significant at the 95% confidence level simply by chance. There is a similar risk that those who conduct ancillary studies, either individually or as a group, may fish for dependent variables until they find one on which the intervention has a positive effect.

In order to prevent the unreliable inferences that come from this type of “fishing,” scholars are advised to disclose all comparisons. In the contexts of ancillary studies of experiments, this requires both comprehensively reviewing other research based on the same intervention and sharing the analysis protocol for the ancillary study.

First, by clearly describing the effects observed in previous studies based on the same intervention, scholars provide readers with information that can help them decide the likelihood the study is measuring a true effect, rather than chance variation. Both the number of previous studies and the substance of their findings are important in making this assessment. For example, questions could sensibly be raised about a job training intervention that was not previously observed to affect employment opportunities but is subsequently found to affect income. Indeed, an important facet of ancillary studies is that we have some priors about the effects of the intervention.

Second, it is important for scholars conducting ancillary studies of experiments to be transparent in their research protocols. Pre-analysis plans are one important mechanism of ensuring greater transparency in research protocols.⁵⁴ However, pre-analysis plans may have other advantages too for ancillary studies. For example, when ancillary studies draw on interventions designed by others, scholars may find that pre-analysis plans are helpful in distinguishing their analysis from that of the original experimenter. These plans also allow the original experimenter to fully assess any risks to the original experiment’s integrity posed by the ancillary study’s research protocol, a key component of experiment sharing that we discuss further below.

4.4 Mechanism Testing

Scholars conducting secondary analyses face particularly great challenges evaluating the causal mechanisms by which the initial treatment affects their

⁵⁴ Pre-analysis plans should be registered before it is possible to begin analysis, which – in the case of ancillary studies – could be before the outcome data is collected *or* before the outcome data has been attached to the data on the initial randomization.

outcome for two reasons. The first is, as in an observational study, they have no control over the experimental design. As a result, they cannot use many of the design-based techniques for identifying causal pathways (Imai et al. 2013). The second impediment to mechanism testing is the time lapse between the original intervention and the new outcomes of interest in the ancillary study. The time lapse often causes the possible mechanisms by which the original intervention could have effects to multiply which makes ruling out rival mechanisms difficult. For example, studies of the effects of an NGO's programming must consider not simply the direct effect of receiving the programme but also any indirect economic or social consequences of the programme that could affect long-term outcomes. Given the increased emphasis in social science on identifying causal mechanisms, this is an important limitation.

Still, mechanism testing is not impossible for ancillary studies of experiments. A number of scholars have assessed the plausibility of competing mechanisms by collecting data on mediating variables and placebo outcomes. For example, Gay (2012) argues that the costs of registering to vote at a new address are unlikely to cause the lower voting rates she observes among individuals who moved out of public housing as part of the Moving to Opportunities program; as evidence, she shows that treated individuals were not less likely to be registered to vote, just less likely to turn out. Similarly, De La O (2013) argues that the positive effect she finds of conditional cash transfers on support for the incumbent is unlikely to be due to clientelism because she does not find any effect of conditional cash transfers on the number of party observers sent to monitor elections. The lack of effects of interventions on intermediary outcomes can help rule out mechanisms.⁵⁵

In another example of mechanism testing, Erikson and Stoker (2011) provide evidence that the Vietnam draft lottery number affected young men's political attitudes toward the Vietnam War by changing their vulnerability to serving in the war using placebo tests. They consider the effect of the 1969 draft lottery on the political opinions of college-bound men in 1973, who would have been able to defer military service during the previous four years but would have been facing imminent military service in 1973 if they had a low draft number. In addition to the college-bound men in their sample whose concerns about serving in Vietnam would have been strong at the time of the survey, they consider the effect of having a low draft number in the 1969 lottery on non-college bound men in 1973 (who would not have been able to defer service and who would either have been drafted or not by this time) and women born on the same birthdates. The fact that

⁵⁵ However, this type of analysis can break down in the presence of heterogeneous effects across sub-groups. For discussions, see Bullock et al. (2010) and Glynn (2012).

they do not find similar effects of draft numbers on these placebo populations allows them to rule out some of the most obvious alternative mechanisms.⁵⁶

Scholars need to do a great deal of work to match previous experiments to unexplored social scientific questions, to collect data on the randomization scheme, and to measure the new outcomes. But as is clear from the large and increasing number of ancillary studies of experiments many scholars have found it feasible to overcome the challenges of ancillary studies to excellent effect. The final section of this essay discusses steps researchers can take to facilitate subsequent ancillary studies while also highlighting the responsibilities of ancillary analysts to maintain the integrity of the original scholar's research design.

5 Best Practices for Experiment Sharing

We believe that the sharing of experiments can benefit both scholars of the original intervention, the Principal Investigators (PIs) who design RCTs, and those conducting ancillary studies. What is the benefit for the scholars of the original study? First, and most obviously, the promise of increased citations. But beyond that, collaboration with scholars conducting ancillary studies can reduce the costs and mitigate the risks of the original scholars. For example, original and ancillary researchers could pool resources, which might permit both sets of scholars to collect more information than either could on their own. There may also be the possibility for original authors to co-author publications with ancillary analysts. So what can be done to facilitate ancillary studies of experiments?

There are a number of steps scholars can take to facilitate the subsequent use of their randomized interventions to identify ancillary effects. As outlined in the previous section, ancillary scholars must be able to identify randomized experiments, gain access to the initial randomization scheme and measure new outcomes over the original experimental units. There are steps scholars can take to facilitate each of these activities.

First, they could register their research designs with organizations such as J-PAL, the Experiments in Politics and Governance (EGAP) network, or the American Economic Association's RCT registry, and they can publicize their results even if they are not statistically significant, activities that are good practice for reasons

⁵⁶ In the future, scholars may also try to repurpose completed experiments to study the effects of mediators on outcomes of interest, as suggested in Ludwig et al. (2011).

of transparency and bias reduction, too.⁵⁷ The registration of experiments helps scholars setting up ancillary studies, since it provides them with centralized databases of experiments from which to start their search. This is particularly useful in flagging studies that are usually hard to find, including ones in-progress, and those that have not been published, perhaps because the original results were not surprising or the effects on the initial outcome were not sufficiently large.⁵⁸

In addition, scholars could consider the potential value of their experiment to future researchers when applying for institutional review board (IRB) clearances. Scholars seeking IRB approval for their research might promise to keep all data confidential in the hopes that this will result in faster approval. But promises to remove all identifiers before publishing the data make the research less valuable to future scholars. In particular, the benefits of the research to the academic community will be greater if the randomization scheme can be shared. Although there are usually strong reasons for both scholars and IRBs to ensure individual-level identifiers are scrubbed from data sets prior to publishing them, when randomization has occurred at the community level, scholars ought to carefully weigh the costs and benefits of promising to remove community-level identifiers before sharing the data. When community-level identifiers can be shared with future scholars, this increases the possibility for future researchers to follow-up on earlier experiments.⁵⁹ At a minimum, scholars will typically need to revise the “off-the-shelf” IRB consent script if they are to maximize the potential for follow-up on their experiments.

Finally, scholars should think carefully about potential future uses of their data when seeking the consent of respondents and tracking compliance. The broader the consent sought and the longer compliance is tracked, the greater the possibility for ancillary analysis.

Ancillary scholars also have a number of responsibilities to the original experimenters. Most obviously, they should cite and prominently acknowledge original studies. Second, ancillary analysts are responsible for ensuring that

57 J-PAL's Hypothesis Registry is available online at: <http://www.povertyactionlab.org/Hypothesis-Registry> EGAP's design registration is available online at: <http://e-gap.org/design-registration/> The American Economic Association's RCT registry is available online at: <https://www.socialscienceregistry.org/>.

58 Null results do not necessarily disqualify an experiment from being of use to ancillary analysts; it is possible for an experiment to be underpowered with respect to identifying effects on the initial outcome of interest but to have sufficient power to identify a different effect (that is expected to be larger). However, null results may also signal problems with the experimental design (i.e., weak prompts, contagion), which warrants caution.

59 This is good practice for non-experimental research too, as it allows the combining of multiple datasets.

their work does not interfere with the initial experimentalists' goals. The original researchers will typically have invested considerable time and resources into their experiment. In order to avoid undermining the original analysis, scholars conducting ancillary studies of experiments should start by informing the original researcher of their proposed research, and sending them a full set of protocols. The two researchers could then assess the risks the second study poses to the initial experimental analysis.

Importantly, if the original researchers are contacted while their data collection is still on-going, they may be open to collaborating with the ancillary analyst to study the second outcome. Collaboration mitigates the risk the original scholar has accepted by investing their time and research funds in the randomized intervention because it provides additional opportunities for publication based on the experiment. Early collaboration also benefits the ancillary study, ensuring the analyst has access to data and protocols from the original experiment. Indeed, when scholars join together early enough, there may be room for the original experimental protocols to be adapted to facilitate study of the outcomes of interest to the ancillary scholar. This breaks down the distinction between RCTs and ancillary studies but is one potential model for experiment sharing. In our own experience, scholars are often receptive to collaborating in this way, so long as the ancillary project is well-specified and does not interfere with the original analysis. If collaboration is out of the question, the ancillary analyst will typically have to wait until the original researchers' data collection is complete before embarking on their project.

The increased possibilities for scholars to collaborate on ancillary studies of experiments could lead to more RCTs in the first place, as scholars consider the benefits of these additional studies when doing their initial cost-benefit calculations. Eventually, it may make sense to establish a formal organization that can manage the sharing of costs and research opportunities provided by large RCTs. Indeed, social scientists engaged in survey and on-line experiments have been sharing space on the same survey platforms through Time-Sharing Experiments for the Social Sciences (TESS) for over a decade now, and this initiative provides a potential model for resource sharing.⁶⁰ But for now, we hope that with good sense and mutual respect, scholars can co-operate to facilitate ancillary studies.

60 In the case of TESS, the principal investigators apply for grants to fund data collection, and then issue competitive calls for proposals. The winning proposals are given space on the TESS survey platform for free. For more information on TESS, see www.tessexperiments.org. For a history of the initiative, see Sniderman (2011).

6 Conclusion

Ancillary studies of experiments are a research method that draws on the merits of both experimental and non-experimental studies. While the method of causal inference in an ancillary study is squarely experimental – insofar as it relies on the randomized assignment of a treatment to make a causal claim – the research tasks involved include the collection of data on the new outcomes being considered, which is an activity more usually associated with observational studies.

Because conducting an ancillary study only requires the collection of observational data, ancillary studies typically have lower research costs than researchers running RCTs. In addition, because the authors of ancillary studies do not bear the responsibility of randomizing the intervention, they are often able to study topics that are ethically or logistically unsuited for RCTs. Ancillary studies draw on found experiments, conducted by other academics for reasons of evaluation or governments for reasons of fairness. As a result, they have been able to study the effects of many large-scale government interventions on sensitive topics.

This study has also noted some of the limitations in the accomplishments of ancillary studies of experiments in economics and political science to date. Although ancillary studies have shown promise in studying some topics related to government performance that are difficult to study using RCTs, the clustering of ancillary studies in certain substantive areas raises concerns about the breadth of this technique's applicability. Indeed, the subjects that can be studied through found experiments will always be circumscribed by what governments, institutions, and researchers are able and willing to randomize. Yet, because researcher-designed RCTs provide one of the types of randomized interventions upon which ancillary studies can build, the substantive areas analysed by ancillary studies should expand with the growth of researcher-designed RCTs.

Acknowledgments: We thank Michael Bernhard, Rajeev Dehejia, Ana De La O, Rachel Gisselquist, Donald Green, Macartan Humphreys, Cindy Kam, Petia Kostadinova, Staffan Lindberg, Fernando Martel García, Miguel Niño-Zarazúa, Elizabeth Levy Paluck, participants at the UNU-WIDER workshop on “Experimental and Non-Experimental Methods in the Study of Government Performance,” three anonymous reviewers and the editors for helpful discussions and feedback, and Sarah Bouchat for superb work on putting together the ancillary studies of experiments database. Thanks also to the numerous scholars who responded to our emails eliciting feedback on the database. A previous essay on this topic was published in *APSA Comparative Democratization* 9/3 (October 2011), and we thank its editors for permission to reproduce parts of that text.

References

- Abrams, D. S. and A. H. Yoon (2007) “The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability,” *University of Chicago Law Review*, 74(4):1145–1177.
- Agarwal, S., S. Chomsisengphet and C. Liu (2010) “The Importance of Adverse Selection in the Credit Card Market: Evidence from Randomized Trials of Credit Card Solicitations,” *Journal of Money, Credit and Banking*, 42(4):743–754.
- Angelucci, M., D. Karlan and J. Zinman (2015) “Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco,” *American Economic Journal: Applied Economics*, 7(1):151–182.
- Angrist, J. D. (1990) “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records,” *American Economic Review*, 80:313–316.
- Angrist, J. D. and S. H. Chen (2008) *Long-Term Economic Consequences of Vietnam-Era Conscription: Schooling, Experience and Earnings*. Discussion Paper 3628. Bonn: IZA.
- Angrist, J. D., E. Bettinger, E. Bloom, E. King and M. Kremer (2002) “Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment,” *American Economic Review*, 92(5):1535–1558.
- Angrist, J. D. and A. B. Krueger (1992) *Estimating the Payoff to Schooling Using the Vietnam-Era Draft Lottery*. Working Paper 4067. Cambridge, MA: NBER.
- Angrist, J. D., S. H. Chen and B. R. Frandsen (2010) “Did Vietnam Veterans Get Sicker in the 1990s? The Complicated Effects of Military Service on Self-Reported Health,” *Journal of Public Economics*, 94:824–837.
- Bagues, M. and B. Esteve-Volart (2011) *Politicians’ Luck of the Draw: Evidence from the Spanish Christmas Lottery*. Working Paper 2011-01. Madrid: FEDEA.
- Bagues, M. and M. J. Perez-Villadoniga (2012) “Do Recruiters Prefer Applicants With Similar Skills? Evidence from a Randomized Natural Experiment,” *Journal of Economic Behavior & Organization*, 82:12–20.
- Baird, S. J. (2007) *Three Seemingly Unrelated Essays in Development Economics*. PhD dissertation. Berkeley: University of California-Berkeley.
- Baird, S., J. H. Hicks, M. Kremer and E. Miguel (2011) *Worms at Work: Long-run Impacts of Child Health Gains*. Working Paper 2011/10. Cambridge, MA: Poverty Action Lab.
- Barnhardt, S. (2009) *Near and Dear? Evaluating the Impact of Neighbor Diversity on Inter-Religious Attitudes*. Job Market Paper 2009/11/10. Cambridge, MA: Harvard University.
- Barrett, C. B. and M. R. Carter (2010) “The Power and Pitfalls of Experiments in Development Economics: Some Non-Random Reflections,” *Applied Economic Perspectives and Policy*, 32(4):515–548.
- Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande and P. Topalova (2009) “Powerful Women: Does Exposure Reduce Bias?” *Quarterly Journal of Economics*, 124:1497–1540.
- Beaman, L., E. Duflo, R. Pande and P. Topalova (2012) “Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India,” *Science*, 335:582–586.
- Bergan, D. E. (2009) “The Draft Lottery and Attitudes Towards the Vietnam War,” *Public Opinion Quarterly*, 73(2):379–384.
- Bhavnani, R. (2009) “Do Electoral Quotas Work after they are Withdrawn? Evidence from a Natural Experiment in India,” *American Political Science Review*, 103(1):23:35.
- Boisjoly, J., G. J. Duncan, M. Kremer, D. M. Levy and J. Eccles (2006) “Empathy or Antipathy? The Consequences of Racially and Socially Diverse Peers on Attitudes,” *American Economic Review*, 96(5):1890–1906.

- Bruhn, M. and D. McKenzie (2009) "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1(4):200–232.
- Bullock, J., D. Green and S. Ha (2010) "Yes, But What's the Mechanism? (Don't Expect an Easy Answer)," *Journal of Personality and Social Psychology*, 98(4):550–558.
- Chetty, R., J. Friedman, N. Hilger, E. Saez, D. Schanzenbach and D. Yagan (2011) "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *The Quarterly Journal of Economics*, 126(4):1593–1660.
- Chattopadhyay, R. and E. Duflo (2004) "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India," *Econometrica*, 72(5):1409–1443.
- Clingingsmith, D., A. I. Khwaja and M. Kremer (2009) "Estimating the Impact of the Hajj: Religion and Tolerance in Islam's Global Gathering," *Quarterly Journal of Economics*, 124(3):1133–1170.
- Conley, D. and J. A. Heerwig (2009) *The Long-Term Effects of Military Conscription on Mortality: Estimates from the Vietnam-Era Draft Lottery*. Working Paper 15105. Cambridge, MA: NBER.
- Cullen, J. B., B. A. Jacob and S. Levitt (2006) "The Effect of School Choice on Participants: Evidence from Randomized Lotteries," *Econometrica*, 74(5):1191–1230.
- Deaton, A. (2010) "Instruments, Randomization and Learning About Development," *Journal of Economic Literature*, 48(2):424–455.
- De La O, A. (2013) "Do Conditional Cash Transfers Affect Electoral Behavior? Evidence from a Randomized Experiment in Mexico," *American Journal of Political Science*, 57(1):1–14.
- De La O, A. and D. Rubenson (2010) "Strategies for Dealing with the Problem of Non-overlapping Units of Assignment and Outcome Measurement in Field Experiments," *The Annals of the American Academy of Political Science*, 628(1):189–199.
- De Paola, M. and V. Scoppa (2011) *Gender Discrimination and Evaluators' Gender: Evidence from the Italian Academy*. Working Paper 06-2011. Consenza: Universita Della Calabria.
- de Walque, D. (2007) "Does Education Affect Smoking Behaviors? Evidence Using the Vietnam Draft as an Instrument for College Education," *Journal of Health Economics*, 26:877–895.
- DiNardo, J. (2008) "Natural Experiments and Quasi-Natural Experiments." In: (S. N. Durlauf and L. E. Blume, eds.) *The New Palgrave Dictionary of Economics*, Second Edition. New York: Palgrave Macmillan.
- Dobkin, C. and R. Shabani (2009) "The Health Effects of Military Service: Evidence from the Vietnam Draft," *Economic Inquiry*, 47(1):69–80.
- Doyle, Jr., J. J., S. M. Ewer and T. H. Wagner (2010) "Returns to Physician Human Capital: Evidence from Patients Randomized to Physician Teams," *Journal of Health Economics*, 29:866–882.
- Duflo, E., R. Glennerster and M. Kremer (2007) "Chapter 61 Using Randomization in Development Economics Research: A Toolkit," *Handbook of Development Economics*, 4:3896–3962.
- Duncan, G. J., J. Boisjoly, M. Kremer, D. M. Levy and J. Eccles (2005) "Peer Effects in Drug Use and Sex Among College Students," *Journal of Abnormal Child Psychology*, 33(3):375–385.
- Dunning, T. (2012) *Natural Experiments in the Social Sciences: A Design-Based Approach*. New York: Cambridge University Press.
- Dynarski, S., J. Hyman and D. Schanzenbach (2013) *Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion*. NBER Working Paper Series No. 17533. Cambridge, MA: National Bureau of Economic Research.

- Eisenberg, D. and B. Rowe (2009) “The Effect of Smoking in Young Adulthood on Smoking Later in Life: Evidence based on the Vietnam Draft Lottery,” *Forum for Health Economics & Policy*, 12(2):1–32.
- Erikson, R. S. and L. Stoker (2011) “Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes,” *American Political Science Review*, 105(2):221–237.
- Ferraz, C. and F. Finan (2008) “Exposing Corrupt Politicians: The Effects of Brazil’s Publicly Released Audits on Electoral Outcomes,” *Quarterly Journal of Economics*, 123(2):703–745.
- Fienberg, S. (1971) “Randomization and Social Affairs: The 1970 Draft Lottery,” *Science*, 171(3968):255–261.
- Frank, D. H. (2007) *As Luck Would Have It: The Effect of the Vietnam Draft Lottery on Long-Term Career Outcomes*. Working Paper, 30 June. Fontainebleau: INSEAD.
- Foster, G. (2006) “It’s Not Your Peers, and It’s Not Your Friends: Some Progress Toward Understanding the Educational Peer Effect Mechanism,” *Journal of Public Economics*, 90:1455–1475.
- Gaines, B. J., T. P. Nokken and C. Groebe (2012) “Is Four Twice as Nice as Two? A Natural Experiment on the Electoral Effects of Legislative Term Length,” *State Politics & Policy Quarterly*, 12(1):43–57.
- Galiani, S., M. A. Rossi and E. Schargrodsky (2011) “Conscription and Crime: Evidence from the Argentine Draft Lottery,” *American Economic Journal: Applied Economic*, 3:119–136.
- Gay, C. (2012) “Moving to Opportunity: The Political Effects of a Housing Mobility Experiment,” *Urban Affairs Review*, 48(2):147–179.
- Gerber, A. (2011) “Field Experiments in Political Science.” In: (J. N. Druckman, D. P. Green, J. H. Kuklinski and A. Lupia, eds.) *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press.
- Gerber, A. and D. Green (2012). *Field Experiments: Design, Analysis and Interpretation*. New York: W.W. Norton & Company, Inc.
- Gibson, J., D. McKenzie and S. Stillman (2009) *The Impacts of International Migration on Remaining Household Members: Omnibus Results from a Migration Lottery Program*. Discussion Paper 20. London: Centre for Research and Analysis of Migration.
- Gibson, J., D. McKenzie and S. Stillman (2010a) *Accounting for Selectivity and Duration-Dependent Heterogeneity When Estimating the Impact of Emigration on Incomes and Poverty in Sending Areas*. Policy Research Working Paper 5268l. Washington, DC: World Bank.
- Gibson, J., D. McKenzie, S. Stillman and H. Rohorua (2010b) *Natural Experiment Evidence on the Effect of Migration on Blood Pressure and Hypertension*. Discussion Paper 24. London: Centre for Research and Analysis of Migration.
- Gibson, J., D. McKenzie and S. Stillman (2011) “What Happens to Diet and Child Health When Migration Splits Households? Evidence from a Migration Lottery Program,” *Food Policy*, 36:7–15.
- Glynn, A. (2012) “The Product and Difference Fallacies for Indirect Effects,” *American Journal of Political Science*, 56(1):257–269.
- Green, D. and A. Gerber (2002) “The Downstream Benefits of Experimentation,” *Political Analysis*, 10(4):394–402.
- Green, D. and A. Gerber (2012) *Field Experiments: Design, Analysis and Interpretation*. New York: W.W. Norton.
- Green, D. and D. Winik (2010) “Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism among Drug Offenders,” *Criminology*, 48:357–387.

- Goldberg, J., M. S. Richards, R. J. Anderson and M. B. Rodin (1991) "Alcohol Consumption in Men Exposed to the Military Draft Lottery: A Natural Experiment," *Journal of Substance Abuse*, 3:307–313.
- Guryan, J., K. Kroft and M. J. Notowidigdo (2009) "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments," *American Economic Journal: Applied Economics*, 1(4):34–68.
- Han, L. and T. Li (2009) "The Gender Difference of Peer Influence in Higher Education," *Economics of Education Review*, 28:129–134.
- Harrison, G. and J. List (2004) "Field Experiments," *Journal of Economic Literature*, 42(4):1009–1055.
- Hastings, J., T. Kane, D. Staiger and J. Weinstein (2007) "The Effect of Randomized School Admissions on Voter Participation," *Journal of Public Economics*, 91:915–937.
- Hearst, N., T. B. Newman and S. Hulley (1986) "Delayed Effects of the Military Draft on Mortality," *New England Journal of Medicine*, 314(10):620–624.
- Heckman, J. and J. Smith (1995) "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9(2):85–110.
- Hemelt, S., K. Roth and W. Eaton (2013) "Elementary School Interventions: Experimental Evidence on Postsecondary Outcomes," *Educational Evaluation and Policy Analysis*, 35:413–436.
- Henderson, J. (2010) *Demobilizing a Generation: The Behavioral Effects of the Vietnam Draft Lottery*. Working paper, 1 September. Berkeley, CA: University of California, Berkeley.
- Hite, N. (2012) *Economic Modernization and the Disruption of Patronage Politics: Experimental Evidence from the Philippines*. PhD dissertation. New Haven: Yale University.
- Ho, D. and K. Imai (2008) "Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: California Alphabet Lottery, 1978–2002," *Public Opinion Quarterly*, 72(2):216–240.
- Hoddinott, J., J. Maluccio, J. Behrman, R. Flores and R. Martorell (2008) "Effect of a Nutrition Intervention During Early Childhood on Economic Productivity in Guatemalan Adults," *The Lancet*, 371:411–416.
- Hoddinott, J., J. Maluccio, J. Behrman, P. Melgar, A. R. Quisumbing, M. Ramirez-Zea, A. Stein, K. Yount and R. Martorell (2013) "Adult Consequences of Growth Failure in Early Childhood," *The American Journal of Clinical Nutrition*, 98:1170–1178.
- Humphreys, M. (2009) *Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogenous Assignment Probabilities*. Columbia University Working Paper.
- Imai, K., D. Tingley and T. Yamamoto (2013) "Experimental Designs for Identifying Causal Mechanisms," *Journal of the Royal Statistical Society, Series A*. 176(1):5–51.
- Imbens, G., D. Rubin and B. Sacerdote (2001) "Estimating the Effect of Unearned Income on Labor Earnings, Savings and Consumption: Evidence from a Survey of Lottery Players," *The American Economic Review*, 91(4):778–794.
- Karlan, D. and J. Zinman (2009) *Expanding Microenterprise Credit Access: Using Randomized Supply Decisions to Estimate the Impacts in Manila*. Yale University Working Paper.
- Kellerman, M. and K. A. Shepsle (2009) "Congressional Careers, Committee Assignments, and Seniority Randomization in the US House of Representatives," *Quarterly Journal of Political Science*, 4:87–101.
- Kling, J. (2006) "Incarceration Length, Employment, and Earnings," *American Economic Review*, 96:863–876.
- Kremer, M. and D. Levy (2008) "Peer Effects and Alcohol Use among College Students," *Journal of Economic Perspectives*, 22(3):189–206.

- Li, H., H. Barnhart, A. Stein and R. Martorell (2003) "Effects of Early Childhood Supplementation on the Educational Achievement of Women," *Pediatrics*, 112(5):1156–1162.
- Lindo, J. M. and C. F. Stoecker (2012) *Drawn into Violence: Evidence on 'What Makes a Criminal' from the Vietnam Draft Lotteries*. Working Paper 17818. Cambridge, MA: NBER.
- Loewen, P. J., R. Koop, J. Settle and J. J. Fowler (2014) "A Natural Experiment in Proposal Power and Electoral Success," *American Journal of Political Science*, 58(1):189–196.
- Ludwig, J., J. Kling and S. Mullainathan (2011) "Mechanism Experiments and Policy Evaluations," *Journal of Economic Perspectives*, 25(3):17–38.
- Maluccio, J., J. Hodinott, J. Behrman, R. Martorell, A. Quisumbing and A. Stein (2009) "The Impact of Improving Nutrition During Early Childhood on Education Among Guatemalan Adults," *The Economic Journal*, 119:734–763.
- Martorell, R., J. R. Behrman, R. Flores and A. D. Stein (2005) "Rationale for a Follow-up Study Focusing on Economic Productivity," *Food Nutrition Bulletin*, 26 (2 Supplement 1):S5–S14.
- McKenzie, D., J. Gibson and S. Stillman (2006) *How Important is Selection? Experimental vs. Non-experimental Measures of the Income Gains from Migration*. Working Paper 06-02. Wellington: Motu Economic and Public Policy Research.
- McKenzie, D., J. Gibson and S. Stillman (2007a) *A Land of Milk and Honey with Streets Paved with Gold: Do Emigrants have Over-Optimistic Expectations about Incomes Abroad?* Discussion Paper. London: Centre for Research and Analysis of Migration.
- McKenzie, David, J. Gibson and S. Stillman (2007b) "Moving to Opportunity, Leaving Behind What? Evaluating the Initial Effects of a Migration Policy on Incomes and Poverty in Source Areas," *New Zealand Economic Papers*, 41(2):197–224.
- Miguel, E. and M. Kremer (2004) "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrics*, 72(1):159–217.
- Ozier, O. (2014) "Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming," Policy Research Working Paper 7052. Washington, D.C.: The World Bank.
- Parker, S. and G. Teruel (2005) "Randomization and Social Program Evaluation: The Case of Progreso," *The Annals of the American Academy of Political and Social Science*, 599:199–219.
- Pollitt, E., K. Gorman, P. Engle, J. Rivera and R. Mortorell (1995) "Nutrition in Early Life and the Fulfillment of Intellectual Potential," *Journal of Nutrition*, 125:111S–118S.
- Rodrik, D. (2009) "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In: (J. Cohen and W. Easterly, eds.) *What Works in Development: Thinking Big and Thinking Small*. Washington, DC: Brookings Institution Press.
- Rohlf, C. (2010) "Does Combat Exposure Make You a More Violent or Criminal Person? Evidence from the Vietnam Draft." *The Journal of Human Resources* 45(2):271–300.
- Rouse, C. E. (1998) "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics*, 113:553–602.
- Rosenzweig, M. R. and K. I. Wolpin (2000) "Natural 'Natural Experiments' in Economics," *Journal of Economic Literature*, 38(4):827–874.
- Sacerdote, B. (2001) "Peer Effects with Random Assignment: Results for Dartmouth Roommates," *Quarterly Journal of Economics*, 116(2):681–704.
- Sekhon, J. and R. Titiunik (2012) "When Natural Experiments are Neither Natural Nor Experiments," *American Political Science Review*, 106(1):35–57.
- Sen, M. (2012) *Is Justice Really Blind? Race and Appellate Review in U.S. Courts*. Working Paper, March 8. Rochester, NY: University of Rochester.
- Siminski, P. and S. Ville (2012) *I Was Only Nineteen, 45 Years Ago: What Can we Learn from Australia's Conscript Lotteries?* Working Paper 12-06. Wollongong: University of Wollongong Economics.

- Sniderman, P. (2011) "The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation." In: (J. N. Druckman, D. P. Green, J. H. Kuklinski and A. Lupia, eds.) *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press.
- Stein, A., M. Wang, A. DiGirolamo, R. Grajeda, U. Ramakrishnan, M. Ramierz-Zea, K. Yount and R. Martorell (2008) "Nutritional Supplementation in Early Childhood, Schooling, and Intellectual Functioning in Adulthood: A Prospective Study in Guatemala," *Archives of Pediatric and Adolescent Medicine*, 162(7):612–618.
- Sondheimer, R. (2011) "Analyzing the Downstream Effects of Randomized Experiments." In: (J. N. Druckman, D. P. Green, J. H. Kuklinski and A. Lupia, eds.) *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press.
- Sondheimer, R. M. and D. P. Green (2010) "Using Experiments to Estimate the Effects of Education on Voter Turnout," *American Journal of Political Science*, 54(1):174–189.
- Stillman, S., D. McKenzie and J. Gibson (2006) *Migration and Mental Health: Evidence from a Natural Experiment*. Working Paper 06-04. Hamilton: University of Waikato Economics.
- Stillman, S., J. Gibson and D. McKenzie (2012) "The Impact of Immigration on Child Health: Experimental Evidence from a Migration Lottery Program," *Economic Inquiry*, 50(1):62–81.
- Stinebrickner, R. and T. R. Stinebrickner (2006) "What Can Be Learned About Peer Effects Using College Roommates? Evidence from New Survey Data and Students from Disadvantaged Backgrounds," *Journal of Public Economics*, 90:1435–1454.
- Stinebrickner, T. R. and R. Stinebrickner (2007) *The Causal Effect of Studying on Academic Performance*. Working Paper 13341. Cambridge, MA: NBER.
- Van Laar, C., S. Levin, S. Sinclair and J. Sidanius (2005) "The Effect of University Roommate Contact on Ethnic Attitudes and Behavior," *Journal of Experimental Social Psychology*, 41:329–345.
- Yakusheva, O., K. Kapinos and M. Weiss (2011) "Peer Effects and the Freshman 15: Evidence from a Natural Experiment," *Economics and Human Biology*, 9:119–132.
- Zinovyeva, N. and M. Bagues (2011) *Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment*. Discussion Paper 5537. Bonn: IZA.