Baldwin, Kate; Bhavnani, Rikhil R.

**Working Paper**

# Ancillary experiments: Opportunities and challenges

WIDER Working Paper, No. 2013/024

**Provided in Cooperation with:**
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

*Suggested Citation:* Baldwin, Kate; Bhavnani, Rikhil R. (2013) : Ancillary experiments: Opportunities and challenges, WIDER Working Paper, No. 2013/024, ISBN 978-92-9230-601-4, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki

This Version is available at:
https://hdl.handle.net/10419/81055



WIDER Working Paper No. 2013/024

# Ancillary experiments

## Opportunities and challenges

Kate Baldwin,[1] and Rikhil R. Bhavnani[2]

March 2013

**Abstract**

'Ancillary experiments' are a new technique whereby researchers use a completed experiment conducted by others to recover causal estimates of a randomized intervention on new outcomes. The method requires pairing new outcome data with randomized treatments the researchers themselves did not oversee. Since ancillary experiments rely on interventions that have already been undertaken, oftentimes by governments, they can provide a low-cost method with which to identify the effects of large-scale and possibly ethically difficult interventions. We define this technique, identify the small but growing universe of studies that employ ancillary experiments in political science and economics, and assess the benefits and limitations of the method.

Keywords: experimental methods, government performance, ancillary experiments, downstream experiments, causal inference, research design
JEL classification: C9, O43, D7



[1]Department of Political Science, University of Florida, email kabaldwin@ufl.edu; [2]Department of Political Science, University of Wisconsin-Madison, email bhavnani@wisc.edu

This study has been prepared within the UNU-WIDER project 'ReCom—Research and Communication on Foreign Aid', directed by Tony Addison and Finn Tarp.

UNU-WIDER gratefully acknowledges specific programme contributions from the governments of Denmark (Ministry of Foreign Affairs, Danida) and Sweden (Swedish International Development Cooperation Agency—Sida) for ReCom. UNU-WIDER also gratefully acknowledges core financial support to its work programme from the governments of Denmark, Finland, Sweden, and the United Kingdom.

ISSN 1798-7237 ISBN 978-92-9230-601-4

**Acknowledgements**

We thank Michael Bernhard, Ana De La O, Rachel Gisselquist, Donald Green, Macartan Humphreys, Cindy Kam, Petia Kostadinova, Staffan Lindberg, Miguel Nino-Zarazua, and Elizabeth Levy Paluck for helpful discussions and feedback, and Sarah Bouchat for superb work on putting together the ancillary experiments database. Thanks also to the numerous scholars who responded to our emails eliciting suggestions for the ancillary experiments 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.



UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Typescript prepared by Anna-Mari Vesterinen at UNU-WIDER.

## 1 Introduction

Field experiments have the potential to advance our social scientific understanding of the world by allowing us to separate cause from effect in natural settings. But this research method also has drawbacks. Field experiments 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 field 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 new data on populations 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 an 'ancillary experiment'. It can be thought of as using 'found' rather than 'designed' experiments.

Ancillary experiments provide many of the advantages of field experiments but at lower cost, since the intervention has already been undertaken. In addition, ancillary experiments can often address questions which are difficult for researcher-designed experiments to study. This is partly because many 'found experiments' are not run by researchers at all. Rather, they are oftentimes lotteries implemented by governments which are less ethically and resource-constrained than individual scholars. As a result, ancillary experiments 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 experiments to analyse these effects. This makes ancillary experiments an attractive tool for scholars studying government performance.

Yet along with the great potential of this type of analysis, this research method has some unique challenges. In this paper, we define and provide an overview of ancillary experiments in economics and political science, and analyse the benefits and limitations of this relatively new research method. We begin by defining ancillary experiments. Next, we take stock of the body of research which uses this technique, drawing on a new, publicly-available database of ancillary experiments.[1] We then provide a discussion of the logistical challenges of conducting this type of research. We conclude by discussing the potential for increased collaboration between scholars to allow the same field experiment to be used to study multiple outcomes.

## 2 Defining ancillary experiments

Ancillary 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 populations randomly assigned to the treatment and control groups in

[1] The database provides a comprehensive listing of ancillary 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 numer of other fields. The database allows us to highlight what this research has accomplished, and also its limitations to date. See, http://rikhilbhavnani.com/research/

the initial intervention. The defining characteristics of ancillary analyses are that they combine new data with the use of a randomized intervention that the researchers themselves did not design or oversee. As a result, ancillary experiments 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.

One of the first examples of this type of research was conducted by Joshua Angrist, who took advantage of the Vietnam draft 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 behaviour and political opinions.[2]

Ancillary experiments 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; Beaman et al. 2012; Bhavnani 2009) and spending (Chattopadhyay and Duflo 2004). A whole new generation of ancillary analyses has been made possible by the increased prevalence of randomized control trials (RCTs) in development economics. These trials do not simply allow the initial researchers to identity 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 graduate students have used a deworming intervention designed and studied by Miguel and Kremer (2004) to study the long-term effects of deworming (Ozier 2010; 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 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 experiments in so far as the researchers took advantage of pre-existing randomized interventions designed and overseen by other scholars or policy makers. We introduce this new term because the phenomenon we are describing encompass a hitherto unrecognized subset of experimental analyses.[3] Ancillary experiments 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 2012; Sekhon and Titiunik 2012).[4] We exclude studies that rely on as-if randomization from our definition, allowing us to focus on the challenges scholars face in using found experiments even when they are explicitly randomized. Ancillary experiments are a broader set of phenomena than 'downstream experiments' as originally conceived by Green and Gerber (2002). As originally defined, downstream experiments use randomized interventions not under the control of the analyst as an instrument to identify the effect of the original outcome on another variable of

[2] Angrist and Chen (2008); Angrist, Chen, and Frandsen (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, Richards, Anderson, and Rodin (1991); Hearst, Newman, and Hulley (1986); Henderson (2010); Lindo and Stoecker (2012); Rohlfs (2005).

3 We thank Don Green for suggesting the terminology. In a previous essay on this topic, we referred to these experiments as 'secondary analyses'. We have changed terminology because the previous term was often conflated with replication studies or meta-analysis.

[4] This is the definition of natural experiments that is currently widely accepted in political science and economics. However, it is worth noting that Harrison and List's earlier definition of natural experiments (2004) also restricted focus to truly randomized treatments.

interest (ibid.: 394). Ancillary experiments include both downstream analyses and analyses that consider the direct effect of the original treatment on new outcomes.[5]

## 3 Taking stock

In this section, we take stock of the use of ancillary experiments to date in economics and political science, drawing on a new database of ancillary experiments. 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.[6] 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 experiments to search for additional studies. Because we found that ancillary experiments often clustered around large randomized interventions, we also searched for articles that mentioned 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 online appendices to the article.

Studies were coded as being ancillary experiments if they used a randomized intervention that the researchers themselves had no role in designing or overseeing *and* they involved new data collection. As a result, we did not count replication studies as ancillary experiments. We also did not include studies that used regression-discontinuity designs or other as-if randomized interventions. To qualify as an ancillary experiment, the study needed to examine populations in which assignment to the treatment and control group was done via an intentionally random process, as is the case in RCTs and lotteries.[7]

Ancillary experiments are distinguished from field experiments 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. In a handful of cases, it was difficult to determine whether or not the scholars played a role in the design of the experiment, and in these cases, we included the papers. It was sometimes also difficult to determine ex post which outcome variables an experiment was initially intended to analyse. We classify studies as ancillary experiments only if they involved new data beyond that collected after the initial field experiment, and if they involved at least one author who was not part of the initial research team. On the one hand, these rules may exclude some ancillary studies in that the same research team may develop new uses for their experiment ex post. On the other hand, they may include some studies where additional data collection and new co-authors were always planned to be brought into the study at a later date. These few misclassifications

[5] 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).

[6] 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').

[7] Some famous lottery studies do not actually meet these criteria. For example, the Imbens, Rubin and Sacerdote (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.

necessarily exist, since we can rarely ascertain exactly what researchers initially intended to do.

As a result of our research, we found 72 studies that qualify as ancillary experiments. Table 1 lists each of these studies, dividing them by whether the treatment involved randomized exposure to a manipulated intervention or randomized exposure to units with different observable characteristics, and the substantive area of the treatment.

Most of the ancillary analyses identified in our search (52 out of 72) take advantage of the randomized exposure of individuals or communities to a specific intervention or programme. For example, randomly selected electoral districts are designated as seats that only women can contest in four studies, and randomly selected individuals receive access to educational programmes in five studies. But a significant number (20 out of 72), use the randomized exposure of individuals (such as judges or roommates) to units with different observable characteristics. The former type of randomization usually results in a stronger design because the treatment to which individuals are exposed is fairly clearly defined. It is much more difficult to isolate the treatment in the case of the latter type of randomization, as the individuals to whom subjects are randomly exposed have multiple attributes.[8] For example, judges who give longer sentences may differ on other dimensions too, making it difficult to be certain that a study examining the impact of having a case assigned to a harsher judge is isolating the effects of long penal sentences. Still, given the randomized assignment of some treatment (even though we do not know precisely what the treatment is), we include these studies in our analysis.

Regarding the substantive themes of the analyses, it is noteworthy that 33 of 72 ancillary experiments in the database relate to government performance, if we define governmental performance studies as those whose dependent 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 55 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 experiments appear better able to do this.

One reason why ancillary experiments are so prevalent in the field of governance is that they build on randomized intervention 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.[9] 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 War, and the Vietnam War,[10] and also purportedly for randomly reserving

[8] Treatments can also be bundled in studies where individuals are randomly exposed to programmes. We elaborate on this concern in the next section.

[9] 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: 208).

[10] 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

electoral seats for female candidates in India. The vast majority of the ancillary experiments we identified rely on lotteries conducted for reasons of fairness which partly explains the prevalence of ancillary experiments in the study of governance. However, it also suggests that RCTs have been largely untapped as a source of ancillary experiments, a fact to which we will return in our critical assessment of the field's accomplishments.

Although ancillary experiments have had some success in studying phenomena—such as large government interventions—that are not easily amendable 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 experiments have been built on just a few *types* of interventions. Even more specifically, there has been a large amount of clustering around specific interventions.[11] For example, 25 per cent of the studies are based on draft lotteries, of which almost 90 per cent use the Vietnam draft lottery. Another 15 per cent of studies are based on international visa/immigration lotteries, of which 90 per cent use the Tonga–New Zealand immigration lottery. Multiple studies have also used the government of India's randomized reservation of seats for women and the Kremer–Miguel deworming experiment (2004) 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 experiments have been applied to date. In considering the strengths and weaknesses of this body of research, it is important to recognize that ancillary 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 five years. Ancillary experiments have just begun to be explored as a research method, and much more can be done with this research technique.

Impartial Random (FAIR) system of selection which will determine the order of call for all equally eligible men.' Quoted in Fienberg (1971).

[11] 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 experiments.

Table 1: Summary of the ancillary experiments database

| Randomized exposure to manipulated intervention or units with different observable characteristics | Substantive area of treatment | Study citations | Number of studies |
|---|---|---|---|
| Randomized exposure to manipulated intervention/ programme | Access to funds/loans | Agarwal, Chomsisengphet, and Liu 2010<br>Bagues and Esteve-Volart 2011<br>De La O 2013<br>Hite 2012 | 4 studies |
| | Military service | Angrist 1990<br>Angrist and Krueger 1992<br>Angrist and Chen 2008<br>Angrist, Chen, and Frandsen 2010<br>Bergan 2009<br>Conley and Heerwig 2009<br>de Walque 2007<br>Dobkin and Shabani 2009<br>Eisenberg and Rowe 2009<br>Erikson and Stoker 2011<br>Frank 2007<br>Gallani, Rossi, and Schargrodsky 2011<br>Goldberg et al. 1991<br>Hearst et al. 1986<br>Henderson 2010<br>Lindo and Stoecker 2012<br>Rohlfs 2005<br>Siminski and Ville 2012 | 18 studies |
| | Educational services | Angrist et al. 2002<br>Cullen, Jacob, and Levitt 2006<br>Hastings et al. 2007<br>Rouse 1998<br>Sondheimer and Green 2010 | 5 studies |
| | Immigration/Visas | Clingingsmith, Khwaja, and Kremer 2009<br>Gibson et al. 2009, 2010, 2011<br>Gibson, McKenzie, Stillman, Rohorua 2010<br>McKenzie et al. 2006, 2007a, 2007b<br>Stillman, McKenzie, and Gibson 2006<br>Stillman, Gibson, and McKenzie 2012 | 10 studies |
| | Housing | Gay 2012 | 1 study |
| | Health services | Doyle, Ewer, and Wagner 2010<br>Baird 2007<br>Baird, Hamory, Kremer, and Miguel 2011<br>Ozier 2010 | 4 studies |
| | Reservation of political seats | Beaman et al. 2009<br>Beaman et al. 2012<br>Bhavnani 2009<br>Chattopadhyay and Duflo 2004 | 4 studies |
| | Political information | Ferraz and Finan 2008 | 1 study |
| | Political power/position | Brockman and Butler 2012<br>Fowler, Koop, Loewen, and Settler forthcoming<br>Gaines, Nokken, and Groebe 2012<br>Ho and Imai 2008<br>Kellerman and Shepsle 2009 | 5 studies |
| Randomized exposure to units with different observable characteristics | Housing/roommates | Barnhardt 2009<br>Boisjoly et al. 2006<br>Han and Li 2009<br>Foster 2006<br>Kremer and Levy 2008<br>Sacerdote 2001<br>Stinebrickner and Stinebrickner 2006<br>Stinebrickner and Stinebrickner 2007<br>Van Laar, Levin, Sinclair, and Sidanius 2005<br>Yakusheva, Kapinos, and Weiss 2011 | 11 studies |
| | Judges | Abrams and Yoon 2007<br>Green and Winik 2010<br>Kling 2006<br>Sen 2012 | 4 studies |
| | Evaluation committees | Bagues and Perez-Villadoniga 2012<br>De Paola and Scoppa 2011<br>Zinovyeva and Bagues 2011 | 3 studies |
| | Workmates | Guryan, Kroft, and Notowidigdo 2009<br>Rogowski and Sinclair 2012 | 2 studies |

Source: ancillary experiments database compiled by authors. See text for details.

### 3.1 What has been accomplished to date

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

The ancillary experiments database demonstrates that ancillary experiments frequently provide a relatively low-cost research technique. First, ancillary experimentalists do not incur any of the costs involved in designing an experiment. Second, although ancillary experiments can involve a wide variety of different data collection techniques (and a wide range of associated costs), in practice, most of the articles and papers in our database collected data on new outcomes from government records or 'off-the-shelf' surveys. Although some of the studies involved expensive follow-up surveys designed by the ancillary researchers, the majority of the studies did not involve the researchers directly interviewing or surveying the populations of interest.

Relatedly, the database shows that it is possible for the same intervention to be used to study a wide variety of outcomes. 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,[12] health outcomes,[13] violence and criminality,[14] and political attitudes.[15] The Tonga-New Zealand migration lottery has been used to study the impact of migration on the income of the migrating family members,[16] the income of those left behind,[17] and the physical and mental health of the migrants.[18] Various roommate studies have analysed the impact of peer effects on inter-racial or inter-religious attitudes,[19] drug and alcohol use,[20] educational outcomes,[21] and weight gain.[22] This indicates possibilities for cost reduction and cost-sharing among scholars interested in a wide variety of substantive outcomes.

The second achievement of ancillary experiments has been their ability to study large-scale interventions and sensitive topics. Many of the found experiments in the database have been implemented by governments. As a result, ancillary experiments provide a useful complement to field experiments, 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, either due to general equilibrium effects, lower capacity, or greater corruption (Deaton 2010; Barrett and Carter 2010). In view

[12] Angrist (1990); Angrist and Krueger (1992); Angrist and Chen (2008); Frank (2007).

[13] Angrist, Chen, and Frandsen (2010); Conley and Heerwig (2009); de Walque (2007); Dobkin and Shabani (2009); Eisenberg and Rowe (2009); Goldberg et al. (1991); Hearst, Newman, and Hully (1986).

[14] Lindo and Stoecker (2012); Rohlfs (2005).

[15] Bergan (2009); Henderson (2010).

[16] McKenzie, Gibson, and Stillman (2006).

[17] Gibson, McKenzie, and Stillman (2009, 2010); McKenzie, Gibson, and Stillman (2007).

[18] Gibson, McKenzie, and Stillman (2011); Gibson, McKenzie, Stillman, and Rohorua (2010); Stillman, Gibson, and McKenzie (2012); Stillman, McKenzie, and Gibson (2006).

[19] Barnhardt (2009); Boisjoly et al. (2006); Van Laar et al. (2005).

[20] Duncan et al. (2005); Kremer and Levy (2008).

[21] Foster (2006); Han and Li (2009); Sacerdote (2001); Stinebrickner and Stinebrickner (2006); Stinebrickner and Stinebrickner (2007).

[22] Yakusheva, Kapinos, and Weiss (2011).

of these concerns, the fact that the majority of ancillary experiments study government-implemented interventions is an advantage. Ancillary experiments can provide important tests of how well programmes scale and are executed by the public sector.

Relatedly, ancillary 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 for 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 ancillary experimentalists 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 conditional cash transfers have been manipulated for another purpose, 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.

The fact that ancillary experiments rely on found experiments; that they do not bear the cost or burden 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 (De La O 2013, Hastings, Kane, Staiger and Weinstein 2007). 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 experiments allow researchers to examine more sensitive and large-scale effects at lower cost than is typically the case with field experiments, ancillary experiments 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 by no means unconstrained. The clustering of ancillary experiments around particular interventions and issues is indicative of such constraints. In this section, we briefly discuss some shortcomings of the corpus of ancillary experiments documented previously.

A striking pattern in our review of ancillary experiments is the scant number of replication studies uncovered. There is a need for greater replication of ancillary experiments 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, 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.[23] Future research should prioritize the replication of ancillary experiments in different settings, through stand-alone follow-up studies or by incorporating results from multiple settings in the initial publication. Publications based on ancillary experiments 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 field experiments. For example, it would be possible for the same scholar to examine the health effects of military drafts in the US, Argentina, and Australia.

Relatedly, this area of research appears to have many randomization-driven searches for questions, but few question-driven searches for randomizations. Of course, it is difficult to determine from final publications 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 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, there have been surprisingly few ancillary experiments building on studies in which the initial randomization was overseen by scholars for reasons of evaluation. Instead, most studies build on interventions that were randomized by governments or other institutions for reasons of fairness. This has provided a useful counter-point 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 experiments 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 a set of three studies conducted by (former) Berkeley graduate students that build on the initial Kremer–Miguel deworming study (Baird 2007; Baird et al. 2011; Ozier 2010). Similarly, Hite (2012) piggybacked on a microfinance experiment run by Karlan and Zinman, and 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, it is surprising that there have not been more ancillary uses of these interventions. The possibility for

[23] 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 (Foster 2006; Han and Li 2009; Sacerdote 2001; Stinebrickner and Stinebrickner 2006; Stinebrickner and Stinebrickner 2007). 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 (Boisjoly et al. 2006; Van Laar et al. 2005), drug use and sex (Duncan et al. 2005), alcohol use (Kremer and Levy 2008), and weight gain (Yakusheva, Kapinos, and Weiss 2011).

collaboration across different sub-fields and even different disciplines in this area is great but largely untapped, a topic to which we will return to at the end of this paper.

## 4 How to conduct an ancillary experiment: major challenges

While ancillary experiments are a new and exciting frontier for research, they are subject to a number of challenges. Some of the challenges of ancillary experiments are shared by experimental design in general (including compliance and spillover problems), and are well-covered in standard texts, including Gerber and Green (2012). Other challenges are shared with natural experiments, although ancillary experiments avoid the 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 experiments 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 conducting an ancillary experiment 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 ancillary experiments 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. This is the first order of business, and demands creativity.

Perhaps the easiest place to start to find a randomized study is the database of ancillary experiments introduced previously. The randomized studies 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. Many existing analyses of experiments have employed lotteries run by governments for reasons for fairness, 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 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 evaluations in economics, provide fairly comprehensive listings of on-going and recently completed field experiments. Many of these field experiments offer opportunities for ancillary experiments, but they also raise questions about norms of experiment-sharing, an issue to which we return in the final section of our study.

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

randomization—Was the lottery carried out properly?[24] How were exceptions dealt with?[25] —and inquiring whether the resulting treatment and control groups are, in fact, balanced in terms of pre-treatment covariates.[26] 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 experiments will typically suggest new pre-treatment covariates on which to check for balance.

In addition, scholars conducting ancillary 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 an ancillary experiment are determined by the original intervention, and not the experimenter. Oftentimes, this means that the population that the ancillary experiment can speak to is narrower than the ancillary experimenter 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.[27]

Care also needs to be taken to understand the degree to which actions in the intervening period affect the original randomization. Experiments involving randomized roll-outs will only rarely be suitable for ancillary experimental analysis.[28] 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).

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

[24] 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.

[25] 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, Ewer, and Wagner (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.

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

[27] 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.

[28] The 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 (2010).

outcome may be anticipated to be smaller or larger 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 and the treatment each unit was actually assigned. 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.

In Bhavnani (2009), this was facilitated by the fact that every electoral district had an equal probability of being selected to be reserved for women. Having documented this, Bhavnani also needed to retrieve information on which electoral districts elected female candidates with and without reservations. The situation is more complicated in randomized evaluations where the probability of receiving treatment varies across different communities, in which case this will need to be documented and accounted for in the analysis.

In some cases, confidentiality agreements employed by the original studies may prevent the primary researcher sharing the randomization scheme. Sharing data may be easier if the scholar conducting the ancillary experiment contacts the individuals responsible for the original study before it is complete. For an in-progress study, Baldwin contacted the scholars conducting the initial experiment early on, and they were able to submit a proposal to the Institutional Review Board (IRB) indicating certain data would be shared with her.

Accessing a randomization scheme is often particularly difficult when the initial treatment is randomized at the individual level for fear of breach of confidentiality. It is noteworthy that in the case of the Vietnam draft lottery, ancillary experimentation has only been possible because randomization was not truly at the individual level. Instead, participants were called by randomly chosen birthdates, information that is more easily obtainable.[29]

Yet many of the ancillary experiments identified in our database do take advantage of individual-level randomizations. In a few studies, the randomization involves a government official being assigned a particular power. In a handful of others, the lottery involves criminal cases being assigned to judges. In both of these instances, there are no confidentiality concerns because of the public status of the units being randomized.

Furthermore, there are a surprising number of studies where scholars are able to recover information on the assignment of private individuals to different treatments from the organizations that ran the lottery. For example, Clingingsmith, Khwaja, and Kremer (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 were 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

[29] Even in this case, scholars have 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).

and to keep the participants' information confidential.[30] 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.[31] Confidentiality concerns make ancillary studies of individual-level randomizations more challenging but not impossible.

Finally, ancillary 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, the average treatment effect on the treated (ATT) may provide a more 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 compliance. In this case, scholars can sometimes generate ATT estimates by using other data sources to estimate the proportion of the treated who take up treatment.[32] 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 correcting estimated effects for multiple comparisons

Another challenge is to measure the outcome(s) of interest in the ancillary experiment. Given the time lag between the original experiment and the ancillary experiment, this often takes significant legwork. For example, in order to conduct their study of 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' (ibid.: 176).

Furthermore, oftentimes the outcome in which the scholar conducting the ancillary experiment 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. 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 are observed line up with each other, which by itself is often a time-intensive undertaking. The problem is that treatment and control units in the ancillary study may not be balanced, because this was not the level at which randomization occurred. 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

[30] Private communication with Donald Green. See also Kremer and Levy (2008).
[31] 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).
[32] See Angrist (1990) for an example.

experimental villages in the precinct. Furthermore, the inclusion of villages not included in the original PROGRESA study potentially created imbalance across the ancillary units.

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 where the ancillary experiment includes populations not included in the initial randomization, researchers must examine whether this creates imbalance and consider the necessity of including additional controls to address this. De La O did this in her study as well, by controlling for the number of villages in each precinct.

The last issue that we wish to raise here is that of making multiple comparisons. Individual studies are increasingly cognizant of the fact that significant effects of interventions are likely to be found with some regularity with the use of many dependent variables. One in 20 dependent variables, for example, are likely to be statistically different from one another in treatment and control groups merely due to chance, where a chance event is defined as one with a 5 per cent probability. The most conservative way to correct for this is by using Bonferroni corrections. Since ancillary studies effectively multiply the dependent variables being considered, albeit across (rather than within) studies, similar corrections should be considered by these authors as well. This type of correction would entail making a complete list of the previous dependent variables considered by studies based on the same intervention, and then adjusting the test statistics used for interpreting the estimated effects in the ancillary experiment. For example, if a scholar decides to examine the effects of the Vietnam draft lottery on a particular outcome after four previous studies have examined different outcomes, the most conservative analysis would consider *p*-values below .01 rather than .05 to be statistically significant at the 95 per cent confidence level.

### 4.4 Mechanism testing

Scholars conducting secondary analyses face particularly great challenges evaluating the causal mechanisms by which the initial treatment affects their 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, Tingley, and Yamamoto 2013). The second impediment to mechanism testing is the time lapse between the original intervention and the new outcomes of interest in the ancillary experiment. 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 experiments. Although it may not always be possible to pin down one causal mechanism using an ancillary experiment, a number of scholars have developed creative ways of successfully eliminating competing mechanisms from consideration. Specifically, scholars have ruled out certain causal pathways through the collection of data on mediating outcomes and the use of placebos. For example, Gay (2012) shows that the costs of registering to vote at a new address are not driving her finding that individuals who move out of public housing are less likely to vote; she does this by demonstrating 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.

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 in 1973, they consider the effect of having a low draft number on non-college bound men whose military fates would have already been decided by 1973, and women born on the same birthdates. The fact that 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.

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 experiments introduced previously, many scholars have found it feasible to overcome the challenges of ancillary experiments detailed above to excellent effect. The final section of this essay discusses steps primary researchers can take to facilitate subsequent ancillary experiments while also highlighting the responsibilities of ancillary analysts to maintain the integrity of the primary scholar's research design.

## 5 Best practices for experiment sharing

Experimental interventions change the histories of treated individuals and communities, allowing scholars to measure multiple different effects of interventions through the observation of these units over time. As a result, just as no individual or organization monopolizes the right to conduct studies in a particular community, no individual or organization monopolizes the right to conduct research that uses the randomization to identify an effect. In order for the academic community to maximize the returns from field experiments, it needs to develop norms of experiment sharing.

There are a number of steps scholars can take to facilitate the subsequent use of their field experiments to identify ancillary effects. They can register their research designs with organizations such as J-PAL or the Experiments in Politics and Governance (EGAP) network, and they can publicize their results even if they are not statistically significant,

activities that are good practice for reasons of transparency and bias reduction, too.[33] The registration of experiments helps scholars setting up ancillary experiments, 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 the ones in-progress, and those that have not been published, perhaps because the primary results were not surprising or the effects on the initial outcome were not sufficiently large.[34]

In addition, primary scholars should consider the potential value of their experiment to future researchers when applying for institutional review board (IRB) clearances, and following up with respondents. For example, scholars generally stop tracking compliance with their interventions once they have finished measuring the primary outcome of interest, however, their experiment will be of greater value to future researchers the longer they document this. In addition, researchers 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 should 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.[35]

Ancillary scholars also have a responsibility to ensure that their analyses do not interfere with the initial experimentalists' goals. The original researchers will typically have investigated considerable time and resources into their experiment. In order to avoid undermining the primary analysis, scholars conducting ancillary experiments should start by informing the primary researcher of their proposed research, and sending them a full set of protocols. It is important for the two researchers to discuss at length any risks the second study poses to the initial experimental analysis.

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. The possibility for future co-authorship with ancillary analysts gives scholars incentives to implement their data collection in a way that facilitates further analysis. Collaboration also allows original and ancillary researchers to pool resources which might permit both sets of scholars to collect more information than either could on their own. 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 primary analysis. If collaboration is out of the question, the ancillary analyst will typically have to wait until the primary researchers' data collection is complete before embarking on their project.

[33] 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/

[34] Null results do not necessarily disqualify an experiment from being of use to ancillary analysts. It may be that the experiment was simply underpowered with respect to identifying effects on the primary outcome of interest. However, null results may also signal problems with the experimental design (i.e., weak prompts, contagion), in which case ancillary experimentalists should be cautious.

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

We believe the possibilities for scholars to collaborate on ancillary experiments could lead to more field experiments 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 to regulate the sharing of experiments. But for now, we hope that with good sense and mutual respect, scholars can co-operate to allow the discipline to learn from the opportunities for ancillary experimentation.

## 6 Conclusion

Ancillary experiments are a new research method that draws on the merits of both experimental and non-experimental studies. While the method of causal inference in an ancillary experiment is squarely experimental—insofar as it relies on the randomized assignment of a treatment to make a causal claim—much of an ancillary experimentalist's research activity involves the collection of data on the new outcomes being considered which is an activity more usually associated with observational studies.

Because the ancillary experimentalists' main activity involves observational data collection, they typically have lower research costs than researchers running RCTs. In addition, because ancillary experimentalists 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 experimentalists look for 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 experiments to date. Although ancillary experiments have shown promise in studying some topics related to government performance that are difficult to study using RCTs, the clustering of ancillary experiments 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 experiments can build, ancillary experiments should be able both to grow with researcher-designed RCTs and to complement its findings in areas less amenable to RCTs.

## 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): 1145-77.

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): 743-54.

Angrist, J.D. (1990). 'Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records'. *American Economic Review*, 80: 313-16.

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): 1535-58.

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: 824-37.

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): 515-48.

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): 379-84.

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): 1890-906.

Broockman, D.E., and D.M. Butler (2012). 'How Do Committee Assignments Facilitate Majority Party Control? Evidence from the Seniority Lottery in the Arkansas State Legislature'. Working Paper. New Haven, CT: Yale University.

Bruhn, M., and D. McKenzie (2009). 'In Pursuit of Balance: Randomization in Practice in Development Field Experiments'. *American Economic Journal: Applied Economics*, 1(4): 200-32.

Chattopadhyay, R., and E. Duflo (2004). 'Women as Policy Makers: Evidence from a Randomized Policy Experiment in India'. *Econometrica*, 72(5): 1409-43.

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): 1133-70.

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): 1191-230.

Deaton, A. (2010). 'Instruments, Randomization and Learning About Development'. *Journal of Economic Literature*, 48(2): 424-55.

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): 189-99.

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 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: 877-95.

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: 866-82.

Duncan, G.J., J. Boisjoly, M. Kremer, D.M. Levy, and J. (2005). 'Peer Effects in Drug Use and Sex Among College Students'. *Journal of Abnormal Child Psychology*, 33(3): 375-85.

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. New York: Cambridge University Press.

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): 221-37.

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): 703-45.

Fienberg, S. (1971). 'Randomization and Social Affairs: The 1970 Draft Lottery'. *Science*, 171(3968): 255-61.

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: 1455-75.

Fowler, J.H., R. Koop, P.J. Loewen, and J. Settler (forthcoming). 'A Natural Experiment in Proposal Power and Electoral Success'. *American Journal of Political Science*.

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: 119-36.

Gay, C. (2012). 'Moving to Opportunity: The Political Effects of a Housing Mobility Experiment'. *Urban Affairs Review*, 48(2): 147-79.

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 (2010). '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, 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.

Gibson, J., D. McKenzie, S. Stillman, and H. Rohorua (2010). 'Natural Experiment Evidence on the Effect of Migration on Blood Pressure and Hypertension'. Discussion Paper 24. London: Centre for Research and Analysis of Migration.

Green, D., and A. Gerber (2002). 'The Downstream Benefits of Experimentation'. *Political Analysis*, 10(4): 394-402.

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: 357-87.

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: 307-13.

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: 129-34.

Harrison, G., and J. List (2004). 'Field Experiments'. *Journal of Economic Literature*, 42(4): 1009-55.

Hastings, J., T. Kane, D. Staiger, J. Weinstein (2007). 'The Effect of Randomized School Admissions on Voter Participation'. *Journal of Public Economics*, 91: 915-37.

Hearst, N., T.B. Newman, and S. Hulley (1986). 'Delayed Effects of the Military Draft on Mortality'. *New England Journal of Medicine*, 314(10): 620-24.

Heckman, J., and J. Smith (1995). 'Assessing the Case for Social Experiments'. *Journal of Economic Perspectives*, 9(2): 85-110.

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): 216-40.

Imai, K., D. Tingley, and T. Yamamoto (2013). 'Experimental Designs for Identifying Causal Mechanisms'. *Journal of the Royal Statistical Society, Series A*. 176 (1).

Imbens, G., D. Rubin, and B. Sacerdotee (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): 778-94.

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*, 96:863-76.

Kremer, M., and D. Levy (2008). 'Peer Effects and Alcohol Use among College Students'. *Journal of Economic Perspectives*, 22(3): 189-206.

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.

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. (2010). 'Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming'. Unpublished.

Parker, S., and G. Teruel (2005). 'Randomization and Social Program Evaluation: The Case of Progresa'. *The Annals of the American Academy of Political and Social Science*, 599: 199-219.

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.

Rogowski, J.C., and B. Sinclair (2012). 'Estimating the Causal Effects of Social Interaction with Endogenous Networks'. *Political Analysis*, 20: 316-28.

Rohlfs, C. (2005). 'Does Military Service Make You a More Violent Person? Evidence from the Vietnam Draft Lottery'. Unpublished.

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.

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 Conscription Lotteries?'. Working Paper 12-06. Wollongong: University of Wollongong Economics

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: 174-89.

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: 1435-54.

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: 329-45.

Yakusheva, O., K. Kapinos, and M. Weiss (2011). 'Peer Effects and the Freshman 15: Evidence from a Natural Experiment'. *Economics and Human Biology*, 9: 119-32.

Zinovyeva, N., and M. Bagues (2011). 'Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment'. Discussion Paper 5537. Bonn: IZA.