# Randomization in field experiments

Péter, Noémi; Soetevent, Adriaan

*Published in:*
Handbook of research methods and applications in experimental economics

*DOI:*
10.4337/9781788110563.00015



*Publication date:*
2019

# 7. Randomization in field experiments

*Noemi Peter and Adriaan R. Soetevent*

## 1 INTRODUCTION

In empirical research, the identification of causal relations furthers our understanding of the processes that underlie human decision-making, economic and social interactions. One of the attractions of experiments is that they can credibly identify the causal effect of a given treatment because researchers can allocate subjects randomly to treatments. Such randomization requires a high level of control. However, as discussed in Chapter 6 in this *Handbook* on advantages and disadvantages of field experiments, ensuring a high level of control is challenging when the experiment is run in the actual field setting of interest (or in a setting that is closer to that than the laboratory). Circumstances in the field are often less malleable and researchers face constraints in the design they can implement and/or the type of data that can be collected. In this chapter, we look at such constraints from the point of view of randomization. We discuss challenges that are posed by these constraints and possible ways to tackle them. We illustrate the potential solutions with examples from articles published in top journals in recent years (2012–17). In particular, we use examples from the *Quarterly Journal of Economics*, the *American Economic Review*, *Econometrica* and the *Review of Economic Studies*.

We lay down the groundwork for the chapter in Section 2, where we review the identification of treatment effects in a formal framework. This section introduces the notation, key concepts and assumptions that will be relevant to the discussion in later sections. It reviews what is meant by potential outcomes, the stable unit treatment value assumption (SUTVA), selection bias and how treatment effects can be estimated in randomized experiments.

In Section 3 we cover three main topics that relate to the level of randomization. In Section 3.1, we focus on potential spillovers in the field that pose challenges to the validity of SUTVA. We discuss various ways to address such challenges, including solutions related to the level of randomization. In Section 3.2, we continue the discussion on the level of randomization and review implementation issues. In Section 3.3, we review the consequences of the level of randomization for the power of the design.

In Section 4, we turn attention to the use of covariates. We explain how the use of covariates in randomization can be advantageous and look at examples of such designs (so-called stratified designs) from the field. We also discuss briefly the use of covariates in the analysis stage. In Section 5, we discuss two ways in which research may fall through. First, we look at what happens if the assignment mechanism is not fully controlled and how this problem can be avoided. Second, we discuss what happens if treatment compliance is only partial. We discuss measures to prevent this problem and possible solutions that can be applied ex post if the problem did occur. Finally, in Section 6 we give a brief summary and provide references for further reading.

## 2 THE IDENTIFICATION OF TREATMENT EFFECTS IN A FORMAL FRAMEWORK

When researchers try to identify treatment effects they have to think about the right comparisons. For instance, imagine that we would like to understand the causal effect of obtaining some specific piece of information on individuals' behavior. To identify this effect, we would need to compare how individuals would behave in the case they obtained the information and in the case they did not. Thus, conceptually we need to think about two different outcomes that could potentially occur. The so-called potential outcome framework can help us do this in a structured way.[1] Here we introduce this framework and have a formal discussion on the identification of treatment effects, leaning heavily on Rubin (1974) and Imbens and Rubin (2015).

We start by reviewing the building blocks. First, Imbens and Rubin (2015, p. 4) define a unit. This may be an individual person or a firm, or a collection of persons, for example a classroom or a market, at a particular point in time. The latter condition carries the important implication that the same person is a different unit at different points in time. Second, units can be exposed to a treatment, also called an action, manipulation or intervention. In the example that we introduced above, the units were individuals and the two alternative treatments were whether the individual obtained the information. As in this example, we will limit attention to settings where there are two alternative treatments. We can then refer to the first one as the active treatment or simply treatment (in our example, receiving the information) and to the second one as the control treatment or simply control (in our example, not receiving the information). The potential outcomes of a unit can be defined as the outcomes associated with the alternative treatments. These outcomes are called "potential" because at any given point in time, only the outcome associated with the administered treatment is observed. For example, if the unit received the active treatment, we can observe only the outcome that is associated with this active treatment, while the outcome that is associated with the control treatment remains an unobservable, counterfactual outcome. The reverse holds for units that received the control treatment. Thus, the relation between the realized (and hence observable) outcome $Y_i^{obs}$ and the potential outcomes can be written as:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) \; if \; W_i = 0 \\ Y_i(1) \; if \; W_i = 1 \end{cases} \tag{7.1}$$

where $W_i$ is an indicator for treatment status with value $W_i = 1$ if unit $i$ receives the active treatment and $W_i = 0$ if unit $i$ receives the control treatment, $Y_i(1)$ is the potential outcome associated with the active treatment and $Y_i(0)$ is the potential outcome associated with the control treatment. An important assumption that underlies equation (7.1) is the stable unit treatment value assumption (SUTVA): "The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms of versions of each treatment level, which lead to different potential outcomes" (Imbens and Rubin, 2015, p. 10).

[1] This notational system is generally referred to as the Rubin causal model, a term coined by Holland (1986), although he readily added (p. 946): "even though Rubin would argue that the ideas behind the model have been around since Fischer."

The first component of SUTVA states that the potential outcomes of a unit are not affected by the treatment status of other units, and hence it suffices to let the potential outcomes depend simply on $W_i$. If this no-interference assumption were violated, we would have to let $Y_i$ depend on the treatment status of other units as well. In particular, one should then write the potential outcome of unit $i$ as $Y_i(\widetilde{\mathrm{W}}^i)$ with $\widetilde{\mathrm{W}}^i$ the $\mathrm{K}^i \times 1$ subvector of the treatment assignment vector W that includes all $\mathrm{K}^i$ units with potential spillovers to unit $i$.[2]

Building on equation (7.1), we can define the effect of the treatment as the difference between $Y_i(1)$ and $Y_i(0)$. As we explained above, one of these two potential outcomes is necessarily unobservable, so the difference between them cannot be calculated.

Instead of the individual treatment effects, researchers often focus on the average treatment effect in the population:[3]

$$\tau_{pop} = E[\,Y_i(1) - Y_i(0)] \tag{7.2}$$

How can we identify this average treatment effect? Let us consider the case where we have information on a random sample of size $N$ from the population of interest. Some units in this sample received the active treatment and others the control treatment. In particular, suppose that the number of units in the active treatment is $N_t = \sum_{i=1}^{N} W_i$, and the number of units in the control treatment is $N_c = \sum_{i=1}^{N} (1 - W_i)$. Then the difference in average outcomes between units that received the active treatment and those that received the control treatment is:

$$\hat{\tau}^{dif} = \overline{Y}_t^{obs} - \overline{Y}_c^{obs} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{obs} - \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{obs} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i(1) - \frac{1}{N_c} \sum_{i:W_i=0} Y_i(0)$$

If the sample is large, this difference approaches:

$$\begin{aligned} &E[\,Y_i(1)|W_i=1] - E[\,Y_i(0)|W_i=0] = \\ &E[\,Y_i(1)|W_i=1] - E[\,Y_i(0)|W_i=1] + E[\,Y_i(0)|W_i=1] - E[\,Y_i(0)|W_i=0] = \\ &E[\,Y_i(1) - Y_i(0)|W_i=1] + E[\,Y_i(0)|W_i=1] - E[\,Y_i(0)|W_i=0] \end{aligned} \tag{7.3}$$

The first term in the last row is the average treatment effect on the treated units. The last term denotes the average outcomes of the units who received the control treatment. The middle term denotes what the average outcome of the treatment group (the actively

---

[2] SUTVA also entails that there are no different versions of the treatment levels. If this assumption were violated – for example, because the active treatment came in two different doses – then we would have to define potential outcomes such that they depend on these different doses.

[3] Imbens and Rubin (2015) distinguish between a finite sample of size $N$ for which one has information and a super-population from which this sample was randomly drawn. Equation (7.2) shows the expected value of the treatment effect in this super-population, which we just call population for simplicity. Note also that by population we actually mean the population of interest. For example, a researcher who is interested in the consumers of a certain product may take a random sample of these consumers. Another researcher might be interested in only those consumers who are between 25 and 40 years old, so he or she may take a random sample from this age group.

treated units) would have been had they received the control treatment. This is an expectation about counterfactual outcomes that are inherently unobservable.

The difference between the middle term and the last term is the so-called "selection bias." It captures the difference between the average outcomes of the treatment and the control group that would have occurred if neither of them received the treatment. In general, we cannot rule out selection bias because the treatment and the control group may be different from each other not only in their treatment but also in many other aspects, such as their motivation, ability, socioeconomic background or some other characteristics that could affect their outcomes. In fact, these differences may be the reason they received different treatments.

Researchers value randomization highly because it can eliminate this selection bias. More precisely, if the treatment status of each unit is random, we expect the treated and the control group to be so similar to each other that their potential outcomes are the same in expectation. That is, proper randomization allows us to assume no selection bias: $E[Y_i(0)|W_i=1] = E[Y_i(0)|W_i=0]$. Moreover, it also means that $E[Y_i(1)|W_i = 1] = E[Y_i(1)|W_i = 0]$, so the average treatment effect on the treated equals the average treatment effect. Thus, equation (7.3) simplifies to:

$$E[Y_i(1)|W_i=1] - E[Y_i(0)|W_i=0] =$$

$$E[Y_i(1) - Y_i(0)|W_i=1] = E[Y_i(1) - Y_i(0)] = \tau_{pop} \tag{7.4}$$

This means that we can credibly estimate the average treatment effect if treatment status is assigned randomly. For example, imagine that we run a randomized experiment on the sample of $N$ units, in which we randomly select $N_t$ units to receive the active treatment, and let the remaining $N_c$ units receive the control treatment. In such a setting, $\hat{\tau}^{dif}$ will be an unbiased estimator of the average treatment effect.[4]

Researchers often use regression techniques to obtain estimates. It is common to specify a linear regression function of the form:

$$Y_i^{obs} = \delta + \tau \bullet W_i + \varepsilon_i, \tag{7.5}$$

with $\tau$ the coefficient of interest, $\delta$ a constant and $\varepsilon_i$ the residual capturing unobserved determinants of the outcome. It can be easily shown that the OLS estimator $\hat{\tau}^{ols}$ is identical to the previous, difference-based estimator:[5]

$$\hat{\tau}^{ols} = \overline{Y}_t^{obs} - \overline{Y}_c^{obs} = \hat{\tau}^{dif} \tag{7.6}$$

Thus, when the OLS estimates are based on data from a well-executed randomized experiment as described above, $\hat{\tau}^{ols}$ will be an unbiased estimator of the average treatment effect.

---

[4] Imbens and Rubin (2015) give a more elaborate, formal proof in Chapter 6 of their book.
[5] See Imbens and Rubin (2015, p. 118).

# 3 LEVEL OF RANDOMIZATION

When a researcher designs a field experiment, one of the most important choices concerns the level of randomization. Should one randomize individuals to different treatments, or should one randomize at a higher level, for example, at the level of the classroom, school, neighborhood or firm? Several aspects need to be considered when making this decision. In Section 3.1, we discuss considerations related to spillovers and identification issues. In Section 3.2, we discuss issues related to implementation. In Section 3.3, we discuss the consequences for the power of the design.

## 3.1 Spillovers and Identification

In field experiments, the physical location and/or the social relationships of the individuals studied are given: individuals take part in the experiment as neighbors, friends, teachers and pupils or employers and employees. This in contrast to laboratory experiments, where subjects enter as singletons and groups and social hierarchies such as leader–follower relationships must be created by the experimenter through role assignment.[6] The essence of many relationships such as those between teachers and pupils are infeasible to create synthetically, so field experiments have a natural advantage in examining certain research questions. However, the fact that the physical location and social context is given poses various challenges as well. In this section, we discuss concerns related to identification.

In Section 2 we relied on the SUTVA when we discussed how the average treatment effect can be identified in a randomized experiment. As we explained, one of the components of SUTVA is that there should be no interference between treatment units. While laboratory experiments typically ensure no interference by preventing certain interactions between subjects, this can be challenging in a field setting. This is especially so when individuals are part of a larger enveloping group, say pupils in a school, neighbors in a street or employees within a firm. A commonly used example of treatment interference (e.g., Baird et al., 2018) is the positive spillovers from vaccination to people who were not vaccinated themselves: the probability of getting a disease drops to zero when all people in your close environment are vaccinated even though you did not receive the vaccination yourself. In this case, a simple comparison of the difference in average outcomes in the control and treatment group gives a biased estimate of the treatment effect. The observed average outcome in the control group is no longer an unbiased estimate of the expected value in the absence of treatment, since units in the control group are in fact affected by the treatment. In the case of positive (negative) spillovers, this will lead to a downward (upward) bias of the average treatment effect. Thus, contamination by treatment inference is an obstacle in estimating the "direct" effect of treatment.

How can a researcher address concerns related to the no-interference part of SUTVA?

[6] Natural bonds such as friendship or study relationships may be present between individuals in standard subject pools, but these ties are often a by-product and not a leading principle in recruiting subjects. Thus, usually only part of a subject's network is observed for a limited range of relations. This limits inference. Experiments with non-standard subject pools may instead utilize existing bonds between subjects by explicitly targeting socially related individuals, for example by bringing employees of the same firm together in an experiment.

If spillovers are likely to occur, researchers should account for this in the design.[7] One can do this by randomizing at the level of the enveloping group within which spillovers are likely to occur. That is, instead of randomizing pupils to different treatments, one can randomize schools to different treatments, so that all pupils who attend a treated school receive the active treatment and all pupils who attend a control school receive the control treatment. Enveloping groups are also called "clusters," and therefore this design is known as a clustered randomized experiment. Note that this design does not eliminate spillovers between individuals. Instead, the idea is to acknowledge that such spillovers are likely to be present, and to aim at estimating treatment effects that internalize these indirect effects. This means that the difference between the outcomes of the treatment and control groups is interpreted as the sum of the direct and the indirect effects.

For example, Fryer (2014) conducts a school-level randomized field experiment to examine how injecting a bundle of best practices from high-performing charter schools into low-performing public schools affects students' performance. The practices are increased instructional time, high-dosage tutoring, data-driven instruction, more effective teachers and administrators and a culture of high expectations. His findings indicate that the treatment has little effect on reading performance, but it significantly increases students' math performance. The effect must be interpreted as a combined effect: a student from a treated school might perform differently than a student from a control school not only because this pupil personally received the treatment, but also partly because his or her schoolmates received the treatment and the resulting change in their behavior might have an additional effect on the pupil. For example, one potential benefit of having a culture of high expectations is that when students try to study hard they are perceived more favorably and are interrupted less by their fellow students.

Fryer's (2014) study highlights another important issue about the level of randomization. While some of the practices would allow randomization at a lower level (such as the level of the individual or the classroom), other practices can only be implemented at the school level. For example, extra tutoring could be randomized at a lower level, while changing the principal can occur only at the school level. Thus, for some treatments, researchers might not have a choice with respect to the level of randomization. We will return to this issue in Section 3.2.

In many cases, researchers are not only interested in the overall treatment effect, but would also like to separately identify the magnitude of the direct treatment effect and of the spillovers. In such cases, a two-step clustered design can help. The first step is to randomize treatment intensity at the cluster level and the second step is to randomize the receipt of the treatment at the lower level. For example, Crépon et al. (2013) set out to examine the direct and indirect effect of job placement assistance on labor market outcomes. Understanding the magnitude of spillovers in this context is important because people may compete for the same jobs. That is, a higher employment rate in the group that received assistance could simply mean that these treated people got some jobs that otherwise would have been filled by people in the control group. Therefore Crépon et al. (2013) conduct a two-step randomized experiment. In the first step, they randomize

[7] Dupas and Miguel (2017, p. 9) contrast the scant empirical attention for estimating treatment externalities in public health and epidemiology to the more embracing attitude of economists who view estimating these externalities as integral to getting at the total treatment effect.

treatment intensity at the level of labor markets: the proportion of job seekers (0, 25, 50, 75, or 100 percent) that will get the treatment is randomly assigned to each city. In the second step, they randomly assign treatment status to job seekers in each city, following the proportions assigned to that particular city. Their results show that treated people are more likely to have found a stable job than people in the control group. However, they also find that the gains may have come partly at the expense of workers who did not benefit from the program.

A simpler version of the two-step design is to have only two treatment statuses at the cluster level. That is, researchers assign a control status to some clusters and a treatment status to other clusters, and then randomly assign some individuals in the treatment clusters to actually receive the intervention. Such a design is applied, for example, by Haushofer and Shapiro (2016) and by Muralidharan and Sundararaman (2015). Haushofer and Shapiro (2016) investigate the response of poor households in rural Kenya to large unconditional cash transfers. Their first step is to randomize villages into treatment and control villages. In the second step, they randomly assign the receipt of cash transfer to households within the treatment villages. This results in three groups of households: "treatment" and "spillover" households in treatment villages, and "pure control" households in control villages. Muralidharan and Sundararaman (2015) investigate how test scores are affected by the provision of private school vouchers. In the first step, they randomize villages into control and treatment groups. In the second step, they conduct a lottery in treatment villages to determine which applicants receive a voucher.

While clustered and two-step designs have the advantage that they can deal with potential spillovers, they also have important drawbacks. It is more difficult to achieve balance when randomizing at a higher level, since the number of enveloping groups is necessarily smaller than the number of individuals. Consider, for example, an experiment in schools. Based on the law of large numbers, randomizing a large number of students into different conditions at the individual level will likely lead to balance between the treatment and the control group. With this we mean that both groups contain a similar number of units that share certain (observed and unobserved) key characteristics. However, the situation is more complicated when randomizing at the school level. There might be only a handful of schools that are willing to participate, and they might be different in many regards. Then the treatment and the control group can never be balanced, no matter which school ends up in which treatment condition. Thus, if we want to randomize at the school level, we must find more schools that are willing to participate.

In addition to the problems with balancing, randomization at a higher level also implies a loss of power. We will discuss power calculations in more detail in Section 3.3. The bottom line is that randomization at a higher level requires a larger sample because of both balancing and power considerations, which can increase the costs substantially.

Given the above-mentioned drawbacks, choosing a higher level of randomization is not necessarily always the best choice. While randomization at a higher level is undoubtedly helpful in the case of spillovers, there may be substantial uncertainty about the probability that spillovers will actually occur. Before conducting the experiment, researchers must think carefully about this. If spillovers are likely to be present, a (two-step) clustered design is recommended. However, if spillovers are deemed to be unlikely, sticking to randomization at the individual level is more cost efficient. Previous research and knowledge of the field can greatly help a researcher in making this decision. Discussing

the issue with partner organizations can also be helpful. Nonetheless, this choice is often a difficult one. An illustrative example is Allcott and Rogers's (2014) study, which examines how households' energy consumption is affected by "Opower reports," that is, by social comparison-based home energy reports. In this study, the experiment was conducted in three different sites as follows:

> In Site 3, which was Opower's first program ever, households were grouped into 952 geographically-contiguous "block batch" groups, each with an average of 88 households, which were randomly assigned to treatment or control. This was done because of initial concern over geographic spillovers: that people would talk with their neighbors about the reports. No evidence of this materialized, and all programs since then, including Sites 1 and 2, have been randomized at the household level. (Allcott and Rogers, 2014, p. 3011)

The nature of the intervention and the way it is implemented can also influence the likelihood of spillovers. When communication between individual units assigned to different treatments (or to the treatment and control group) cannot be ruled out, it can be helpful to take measures that obfuscate that units are exposed to different treatments (or no treatment). For example, when studying the impact of providing certain information, one may send people in the control group a "placebo" mail. Then, even if people in the treatment and control group communicate, they are less likely to find out the treatment variable because both groups have received a letter. However, Dupas and Miguel (2017) notice that in many health-related settings it is almost impossible to blind individuals to their treatment status by providing a placebo treatment.[8]

Another way to reduce the chance of spillovers is to aim for a shorter time frame. If the effect of the treatment can be expected to materialize quickly, it is useful to collect information on outcome variables shortly after the treatment was provided, before there is time for interactions between subjects.

It is important to think about ways to test the no-interference part of SUTVA when randomization occurs at the individual level. An interesting example is provided by Bloom et al. (2015), who examine how working from home affects workers' performance. They randomize volunteer employees of a certain call center into two different treatments: work from home or work in the office, for nine months. They find that those who were assigned to work from home have a substantially higher performance. To investigate whether the results are led by spillovers, they compare the performance of the control group to the performance of comparable workers from another call center of the same company. They conclude that spillovers are unlikely, since they do not find a lower performance in the control group, even though they have lost the treatment lottery.

The no-interference assumption can also be addressed by looking at spatial externalities. This method is used by Callen and Long (2015), who conduct a field experiment in Afghanistan on election fraud during a parliamentary election. They make use of the so-called "photo quick count" technology, which means photographing provisional vote tally sheets at individual polling centers before the aggregation of votes, so that these counts can be compared to the corresponding numbers after aggregation is completed. The aim

---

[8] Dupas and Miguel (2017, p. 7): "To illustrate, how would you provide a placebo treatment in a study investigating the impact of the distribution of antimalarial bed nets?"

of their experiment is to test whether announcing photo quick count measurements via a letter to polling center managers reduces aggregation fraud. It is particularly interesting to look at spatial externalities in this setting as they can operate in two different directions. On the one hand, fraud might be displaced from treated centers to control centers, which creates an upward bias in estimating the treatment effect. On the other hand, fraud might be reduced in polling centers that are close to treated centers if treated managers notify their nearby colleagues of the possibility of photo quick count measurements. This would lead to an underestimation of the treatment effect. To assess these possibilities, Callen and Long (2015) compare votes in control polling centers that have no treated neighbors within a certain radius to votes in control polling centers that do have at least some treated neighbors. They find that the average number of votes for the most connected candidates in the first group is significantly larger than in the remaining controls, which suggests that negative treatment externalities are present.

Another example of analyzing spatial spillovers is provided by Dupas (2014), who examines how short-run subsidies affect the long-run adoption of new health products. In this study, the treatment group receives a subsidy on the price of improved antimalarial bed nets. Dupas (2014) examines not only how the subsidy affects short-run purchases, but also the willingness to buy the product again a year later at a higher price. The treatment variation in Dupas (2014) is different amounts of subsidies on the price of the product, in the range of 40 percent to 100 percent. Thus, all households in the sample are assigned to a subsidized price, but there is variation in the amount of the subsidy. Dupas (2014) conducts the experiment in several areas, and in each of them there are four or five different subsidy levels. She finds that initial take-up is very sensitive to price: the take-up rate is 97.5 percent for free vouchers, but drops gradually with price, and is below 10 percent when the subsidy is only 40 percent. Given the random assignment of households to price groups, there is exogenous variation not only in own treatment status, but also in geographic proximity to other households with improved bed nets. Those who live close to a household with a high subsidy have a different exposure than those who live further from such households, and Dupas (2014) exploits this to examine spillovers. Interestingly, her results indicate a positive spillover in the short run but a negative spillover in the long run. A possible explanation for this is that people react to health spillovers over time. Initially, people get convinced to use a bed net if they see their neighbors using it effectively. However, as malaria rates decrease over time with bed net coverage, having more neighbors who use the bed net leads to a lower private return and hence lowers purchases in the long run.

So far, we focused on the complications that interactions between units create for obtaining an unbiased estimation of causal relations. However, in other cases, the primary object of interest is exactly the effect of interactions between individuals. This is well illustrated by papers that aim to identify peer effects. For example, Booij, Leuven and Oosterbeek (2017) examine peer effects originating from the ability composition of tutorial groups for undergraduate economics students. They manipulate the ability composition of groups by assigning students randomly to tutorial groups, conditional on prior ability. Their results imply that low- and medium-ability students gain from being grouped with peers with similar ability, whereas high-ability students do not.

Bursztyn et al. (2014) aim to identify two types of peer effects in financial decisions. One channel is called "social learning," which means that when someone buys an asset, his or

her peers also want to buy it because they learn from his or her choice. The other channel is "social utility": the fact that an individual owns an asset might directly affect others' utility of owning the asset. To investigate these channels, Bursztyn et al. (2014) set up an experiment in collaboration with a large financial brokerage in Brazil. First, they use previous information on referrals to identify pairs of investors who are socially related. Then they randomly assign the roles of Investor 1 and 2 within each pair. First, Investor 1 gets the opportunity to purchase the asset. If he chooses to purchase it, a lottery is conducted to determine whether the purchase is successful. Then Investor 2 is randomized into one of the following conditions: (a) she receives no information about Investor 1, or (b) she is informed of Investor 1's purchase decision and the result of the randomization that determined possession. By comparing the Investor 2 outcomes in the no-information treatment to the outcomes of those in the information treatment whose paired Investor 1 managed to purchase the asset, Bursztyn et al. (2014) are able to identify the effect of learning plus possession. By comparing the Investor 2 outcomes in the no-information treatment to the outcomes of those in the information treatment whose paired Investor 1 did not manage to purchase the asset, they can identify the effect of learning alone.

A key identification assumption in Bursztyn et al. (2014) is that the no-information group is truly a group that did not learn about the purchase desire and possession of Investor 1. However, the social connection between the two investors creates a risk of communication. To lower this risk, Bursztyn et al. (2014) instruct brokers to call Investor 2 on the same day as they called Investor 1. By making the second call soon after the first, they managed to keep the contamination rate as low as six out of 150 cases. As this example illustrates, a careful design can enable researchers to tackle challenges created by social connections, even when these connections are so strong that they form the basis of the research question.

### 3.2 Implementation Issues

Implementation and practical issues can also affect the choice of the level of randomization. It might be that outcome data is only available at a more aggregate level, so there is no advantage to randomizing at lower levels. For example, in their study on how information about candidates affects voting behavior, Kendall, Nannicini and Trebbi (2015) randomize their information treatments at the precinct level as that is the smallest electoral unit for which outcome data is available.

In other instances, the research question or context might imply that some treatments only make sense at a higher level. When studying household decision-making over fertility, access to contraceptives should be randomized at the household level (Ashraf, Field and Lee, 2014). When employees carry out interconnected tasks and rely on each other heavily, it makes sense to provide bonuses at a team level (Friebel et al., 2017). When evaluating the impact of monitoring on teacher absenteeism by tracking school openings and closings using cameras (Duflo, Hanna and Ryan, 2012), randomization can be implemented only at the school level.

Another consideration is that carrying out randomization at the lower level can be rather difficult. Whereas in laboratory experiments subjects travel to the laboratory to receive treatment, in field experiments treatments are brought to the natural habitat of the subjects. This entails that researchers often need to delegate treatment administration to

research assistants or to their outside partners. To reduce the risk of incorrect treatment administration, it helps to use a level of randomization that delineates treatment groups using a characteristic that is salient to the assistant or outside partner. This keeps the cognitive costs of switching between treatments low. For example, in their door-to-door fundraising experiment on social pressure and altruism, DellaVigna, List and Malmendier (2012) use 92 solicitors and surveyors and randomize the distribution of flyers with treatment-specific information at the street level. One can imagine that the risk of treatment contamination would be higher when randomizing at the level of individual homes. Also, outside partners may block lower-level randomization because of ethical concerns. Even when feasible, schools may deem within-class differences in treatment assignment "unfair."

Finally, in field experiments there is often a trade-off between the two objectives of avoiding treatment interference and creating treatment and control groups that are similar in terms of observable and unobservable characteristics. We touched upon this issue in the previous subsection where we discussed the difficulties of randomizing an education intervention at the school level instead of the student level. In that example, treatment interference was a threat because of the social context. In other instances, challenges arise because the fixed physical location of participants carries a risk of interactions. For example, if two neighbors are subjected to distinct treatments, they may discuss these and adjust their behavior to this treatment variation. On the other hand, if treatments are varied across neighborhoods only, then differences other than those induced by the treatment may affect choices. In contrast, in laboratory experiments one can randomize conditions within sessions without increasing the risk of interaction between subjects seated in different cubicles.

### 3.3 Power Considerations

Randomization at a higher level may reduce treatment interference by internalizing spillovers, but this comes at the cost of reducing the power of the experiment. As a result, a clustered approach may need to be combined with larger samples in order to maintain power. Statistical power calculations answer the question of which sample size is needed to identify a treatment effect of a certain size. Nowadays, researchers routinely perform such calculations as one of the first steps in setting up experiments.[9] They do this with good reason, because getting the sample size right saves time and money: an underpowered design (one with insufficient units) leads to inconclusive results, while in an overpowered design resources are wasted because too many units are sampled. We therefore elaborate in this section on power calculations, relying heavily on Section 4 of Duflo, Glennerster and Kremer (2007), which also builds on Bloom (1995, 2005).

Suppose that we wish to test the null hypothesis of no treatment effect, that is, $H_0$: $\tau = 0$ in (7.5). Then first we decide which probability $\alpha$ of finding false positives we deem acceptable – that is, the probability of rejecting the null when there is no effect (type I error). The common choice is to choose $\alpha = 0.05$; this is the significance level or "size" of

[9] Various packages for statistical power calculations are available – for example, G*Power (Faul et al., 2007; see also http://gpower.hhu.de/, accessed February 8, 2019), Optimal Design (Raudenbusch et al., 2011), or the "power" command in Stata.

the test. The other false inference that can be made is to fail to find an effect (so to fail to reject the null) when there is one (so the null is in fact not true). Ceteris paribus, one aims to maximize the probability $\kappa$ of *not* finding such false negatives (type II errors). This probability $\kappa$ is called the power of the test.[10] Usually, the aim is to achieve a power of at least $\kappa = 0.80$. That is, we want to detect the effect with at least an 80 percent probability if the treatment actually works.

The interesting question then is how large the actual treatment effect $\tau$ needs to be in order to be detected for given values of $\alpha$ and $\kappa$ and a given sample size $N$. This question can be answered by considering the regression function in (7.5). Assuming that the observations $Y_i^{obs}$ are independent and identically distributed (i.i.d), the variance of the OLS estimator $\hat{\tau}^{ols}$ of $\pi$ is given by:

$$var(\hat{\tau}^{ols}) = \frac{\sigma^2}{p(1-p)N} \tag{7.7}$$

with $\sigma^2$ denoting the variance of the residual $\varepsilon_i$ and $p$ the probability of assignment to the treatment group. Suppose that the true effect size is $\tau^{TRUE}$. It can be shown that a test of size $\alpha$ and power $\kappa$ will detect this effect if and only if:

$$\tau^{TRUE} > (t_{1-\kappa} + t_{\alpha})\sqrt{var(\hat{\tau}^{ols})} \tag{7.8}$$

where $t_{\alpha}$ is taken from a standard $t$-distribution.[11] The minimum detectable effect (MDE) size is then given by:

$$MDE = (t_{1-\kappa} + t_{\alpha})\sqrt{\frac{\sigma^2}{p(1-p)N}} \tag{7.9}$$

in the case that the test is one-sided (one should divide $\alpha$ by two in the case of a two-sided test). This expression shows that the minimum effect that one can expect to detect is decreasing in sample size $N$. Also, the MDE is smallest when an equivalent proportion of units is assigned to the treatment and control group, that is, $p = 1/2$. This means that whenever there are practical reasons to administer treatment to deviate from a balanced treatment assignment – for example, because it is too costly to assign half of the eligible persons to a training program – one must bear in mind that this comes at the cost of increasing the MDE.[12] These considerations change when multiple alternatives are tested against one control treatment. In such cases, researchers need to assign more units to the control group than to each of the treatment groups in order to effectively decrease the MDE of each treatment (Duflo et al., 2007, pp. 3920–21 elaborate on this trade-off).

In the design stage, the reverse question to the above is of even higher interest: given that we aim to detect a treatment effect that is at least of size $\tau^{MIN}$, how large should our

---

[10] $P$(type II error) = 1 –$\kappa$.

[11] We refer to Section 4 of Duflo et al. (2007) for a graphical derivation.

[12] It is also common to use the Standardized Effect Size (SES), which is the MDE divided by $\sigma$. For example, SES = 0.25 means a minimum impact equal to 0.25 standard deviations. Common practice is to set up a design with SES = 0.20; designs with SES > 0.50 (only treatment effects larger than half of a standard deviation will be detected) are likely to be underpowered. In papers, it does not seem common to report the power of the design, one exception being Voors et al. (2012).

sample be (given our choices for $\alpha$ and $\kappa$)? The answer is obtained by replacing MDE with $\tau^{MIN}$ in equation (7.9) and solving for $N$:

$$N^* = \frac{(t_{1-\kappa} + t_\alpha)^2 \sigma^2}{(\tau^{MIN})^2 p(1-p)} \tag{7.10}$$

The outcomes of power calculations should be used only as suggestions for what sample size might be appropriate. The usefulness of these calculations relies on the plausibility of the assumptions on which they are based and these assumptions in practice involve considerable guesswork. In particular, one has to provide in advance an estimate of the sample variance of the outcome variable and an assessment of what treatment effect would be realistic. To do this, one can consult empirical estimates in similar studies or run a small pilot study.

In the case of a clustered design, the power of the experiment is lower. First, outcomes might only be available at the group level, which implies a lower sample size. Second, even if one can use individual-level data for the analysis, power will be reduced because the observations are not independent. For example, common shocks at the group level introduce intra-cluster correlation. This decreases power for a given MDE because it reduces the precision of the treatment estimates. In particular, suppose that there are $J$ clusters of identical size $n$ and the regression equation is modified such that it can be written as:

$$Y_i^{obs} = \delta + \tau \bullet W_i + v_j + \omega_{ij} \tag{7.11}$$

with $j$ indexing the group and $v_j$ representing shocks at the group level.[13] If $v_j$ is i.i.d with variance $\gamma^2$ and $\omega_{ij}$ is i.i.d with variance $\sigma^2$ then the new OLS estimator will be unbiased with variance:

$$var(\hat{\tau}_{clust}) = \frac{n\gamma^2 + \sigma^2}{p(1-p)nJ} \tag{7.12}$$

In contrast, if the randomization occurred at the individual level, the variance of the OLS estimator would have been:

$$var(\hat{\tau}_{indiv}) = \frac{\gamma^2 + \sigma^2}{p(1-p)nJ} \tag{7.13}$$

Comparing equations (7.12) and (7.13) makes it clear that randomization at the cluster level indeed increases the variance of the estimator.

## 4 HOW TO OPTIMALLY USE COVARIATES?

At different stages of the experimental project, information on covariates or control variables can be useful in increasing the precision of the estimates. This section discusses the added value of covariates in the design stage and in the analysis stage of the experiment. Ideally, one should give considerable thought to the question on which covariates

[13] This example and the derivation is presented by Duflo et al. (2007, pp. 3921–2), following Bloom (2005).

to collect data prior to the implementation of the experiment, even when one plans to use them only in the analysis stage. One particular covariate of interest is pre-treatment values of the outcome variable.

In the past, running baseline surveys was often the method of choice to collect such baseline information. Nowadays, field partners such as companies increasingly collect and store data on their and their employees' performance as part of their daily operations, providing a rich additional source of baseline data. However, researchers still must pay considerable effort (and be patient) to get access to these data. They can advance their case by presenting a convincing story on how having information on specific covariates will increase the precision of the treatment estimation.[14]

### 4.1 Use of Covariates in Treatment Assignment

Until now, we have implicitly focused on experiments conditioned only on the number of units $N_t$ that one wishes to assign to the treatment group. In other words, experiments for which the assignment vector $\mathbf{W}$ is randomly drawn from the set:

$$\left\{\mathbf{W} \in \mathbb{W}: \sum_{i=1}^{N} W_i = N_t\right\},$$

with $\mathbb{W}=\mathbf{\{0,1\}^N}$ the set of all possible values. Experiments with this assignment mechanism are called completely randomized experiments (Imbens and Rubin, 2015). In Section 2, we discussed that randomization is valued because it leads to balanced treatment and control groups in expectation. However, this does not necessarily mean that it will actually lead to balance *in practice*.[15] If in advance of the experiment information on specific attributes of the units is available, one can use it in the design stage to ensure balance in practice, at least with respect to those attributes. This reduces the sampling variability of the estimated treatment effect. For example, in field experiments within firms, the age and gender of employees may predict the effect they experience from treatment. If so, one would like to assign equal fractions of each gender and age group to the treatment and control group.

Experiments that use block or strata in treatment assignment are called stratified randomized experiments.[16] The first step of a stratified design is to define blocks. Within blocks, units are similar with respect to some covariates or pre-treatment variables that are thought to be predictive of outcomes. Stratification based on gender is a common example. More generally, suppose that based on the covariates a total of $J$ blocks are defined with $B_i \in \{1, \ldots, J\}$ indicating the block of the $i$-th unit. Block $j$ has a total of $N(j)$ members. In a second step, a completely randomized experiment is applied within each block, with the assignment being independent across blocks. If the researcher

---

[14] Relatedly, Carneiro, Lee and Wilhelm (2017) consider the problem of a researcher with a limited budget who has to decide on whether to use the money to expand the number of observations or to expand the number of covariates on which information is available. They develop a procedure to optimally make this trade-off, assuming the researcher's objective is to obtain an accurate estimate of the average treatment effect.

[15] This issue diminishes as the sample size $N$ increases.

[16] Completely randomized experiments and stratified randomized experiments both belong to the class of classical randomized experiments. Two other members are the Bernoulli trial and the paired randomized experiment; see Imbens and Rubin (2015, Chapter 4).

ordains that $N_t(j)$ units in block $j$ are assigned the treatment, the assignment vector $\mathbf{W}$ in a stratified randomized experiment is randomly drawn from the set:

$$\left\{\mathbf{W} \in \mathbb{W}: \sum_{i:B_i=j}^{N} W_i = N_t(j) \text{ for } j = 1, 2, \ldots, J\right\},$$

with $N(j) > N_t(j) > 0$ for all $j = 1,2,...,J$.

If the researcher has access to pre-treatment variables that are thought to be predictive of outcomes, defining blocks based on these variables ensures that the presence of each type is balanced across treatment and control groups. One avoids, for example, the over- or underrepresentation of, say, males in the treatment group. This improved similarity of treatment and control group on pre-treatment variables increases the precision of any causal inference. Even when the blocking indicator $B_i$ has no predictive power, a stratified design does not worsen the actual precision.[17] Thus, in the design phase of any experiment, it pays off to put some effort into collecting information on relevant covariates to enable a stratified design. This is especially so when the level of randomization is at higher levels. As discussed before, the fact that the number of clusters is lower implies that it is more difficult to achieve balance. Therefore, stratification is particularly important in these cases.

Many of the studies we discussed in Section 3 use stratification. Fryer (2014) used a matched-pair randomization procedure, which can be viewed as an extreme version of stratification. First, he ordered the 18 schools in the sample by the sum of their mean reading and math scores from the preceding year. Then he formed "matched pairs" such that the first and the second school on the list constituted the first pair, the third and the fourth school the second pair, and so on. Then from each pair, he randomly assigned one school into the treatment group and the other one into the control group. In their study on job placement assistance, Crépon et al. (2013) used a similar method. They first combined public unemployment agencies that covered areas of comparable size and population into groups of five. Then they randomly assigned the proportion of job seekers that will receive treatment (0, 25, 50, 75, or 100 percent) within each of these groups. In Muralidharan and Sundararaman (2015), the randomization of villages to school voucher programs was stratified by district. In Callen and Long's study (2015) on election fraud, polling centers were randomized to the treatment and control condition after stratification by province and by covariates from a baseline survey.

Finally, we would like to reflect on the fact that so far we talked about a stratified design only as being advantageous. Are there situations in which the researcher might not want to stratify, even if information is available? Concerns about the reactions of participants come to mind at this point. For example, a researcher might be able to observe the gender of participants and could therefore stratify based on that. However, if randomization occurs in a transparent way in front of the participants, they might find stratification strange or worrisome. Why does the researcher want women to pick the envelope, lottery

[17] See Athey and Imbens (2017, Sections 7.2–7.3). They also consider the question whether researchers should engage in re-randomization when one notices after random assignment – but before implementation – that for one or more key covariates there is substantial imbalance between the treatment and control group. Their simple advice is to rule out such imbalance from the outset because ex post corrections complicate inference.

ticket or card from a different pile than men do? Such practices can create suspicions among participants so it may be better to avoid them.

### 4.2 Use of Covariates in Estimation

In addition to using the variable $B$ for stratification, it can also be added as a control variable to equation (7.5) to reduce the residual variance. When the share of treated units is the same within each block, researchers can leave out $B$ from the regression, since the point estimate of $\hat{\tau}^{ols}$ will be unbiased either way. However, in some cases, the share of treated units might differ between blocks. For example, a partner organization might insist on assigning a larger fraction of disadvantaged individuals to a treatment that is believed to benefit them. In such cases, it is essential to control for disadvantaged status in the regression, since assignment is random only conditional on this variable.

Adding other covariates to the regression function in the analysis stage may also prove useful. Under random assignment, the estimator $\hat{\tau}^{ols}$ in equation (7.6) is an unbiased estimator of the treatment effect without a need for adding control variables in the regression equation (7.5). However, the addition of control variables may reduce residual noise and thereby increase the statistical precision of the estimated causal effect. The expression for the minimum detectable effect in equation (7.9) reflects this: a lower residual variance $\sigma^2$ decreases the MDE and thereby increases the power of the statistical test. The gain in precision increases the more the covariates help in predicting the potential outcomes and is zero when the covariates are in fact unrelated to the potential outcomes.[18]

It is good practice to report the results of a regression that only includes the treatment indicator $W_i$ next to results that include additional covariates: under random assignment, both provide an unbiased estimate of the true treatment effect such that finding no significant differences in estimated effects allays concerns that treatment assignment has been non-random.

## 5 THE DEVIL FOOLS WITH THE BEST LAID PLAN

No matter how much thought one has given to the design of an experiment, unexpected things are sure to happen in practice. Some of these eventualities can be anticipated to a certain extent, however. This section discusses a number of them but the list is non-exhaustive.

### 5.1 What if the Assignment Mechanism is Not Fully Controlled?

In the ideal case, the researcher has perfect control over the assignment mechanism and the randomization procedure as laid down in the experimental design is implemented. However, circumstances in the field often deviate from this ideal setting and the planned assignment mechanism may not be followed through in practice.

---

[18] See Imbens and Rubin (2015, Section 7.8). They point out that adding covariates will not lower precision in sufficiently large samples. In small to moderate samples, it may lead to a loss in precision due to a loss in degrees of freedom. In such cases, it is advisable to limit the number of covariates to a relatively small number that is suggested by economic rationale.

We distinguish between two cases. In the first case, unforeseen circumstances prohibit the researcher from implementing the assignment as planned. For example, one plans to administer different treatments to a certain number of units but not all units show up or an insufficient number of research assistants prevents treatment administration to all units available for treatment.[19] In such cases, it is often possible to arrange the administration in such a way that randomization is not compromised despite the sample being smaller.

In the second case, the researcher faces uncertainty on whether the assignment has been truly random, for example because treatment administration has been done by assistants or by employees of the collaborating partner. In those cases, treatment mix-ups are likely to increase with the distance between the researcher and these employees. We advise keeping this distance as small as possible.[20] In the design stage, specific instructions can help to alleviate problems in implementation. Instead of saying that the assignment should be random, the researcher could specify which randomization mechanism to use (e.g., flip a coin, use a deck of cards, conduct a lottery). Otherwise, employees will use their own assignment rules that might be biased. For example, when employees are simply instructed to assign the treatment randomly to half of the people who show up, they might assign the treatment to everyone who shows up early, and to no one who shows up later. It is conceivable that this can result in an imbalance between the treated and untreated group. In the analysis stage, a comparison of the covariate distributions of the treatment and control can shed light on the severity of the imbalance.

### 5.2 What if Treatment Compliance is Only Partial?

Non-compliance means that participants get a different treatment than the one they were assigned to. It may be that some participants in the treatment group go untreated but also that members of the control group unintendedly receive the treatment. In both cases, comparing the treated with the untreated will lead to a biased assessment of the treatment impact because the group that is *actually* treated is a (self-)selected, thus non-random subsample. There are several ways in which non-compliance can be prevented or reduced in the design stage. First, researchers can try to increase take-up among those who were assigned to the treatment by making take-up easy and/or attractive. This can be done, for example, by providing incentives or non-pecuniary encouragements, like putting the text "Important – Good News for You" on envelopes (Bhargava and Manoli, 2015). Second, placebo treatments or strict monitoring can help to avoid take-up of the active treatment by the control group. Third, as discussed in Section 5.1, researchers should aim to stay close to those actually administering the treatment.

---

[19] This happened to Soetevent in a door-to-door fund-raising experiment with randomization at the solicitor level (Fosgaard and Soetevent, 2018). Many solicitors showed up at around the same time such that the research assistants could not instruct all of them. As a result, the number of solicitors in the sample was smaller than initially planned.

[20] For this reason, in the door-to-door fund-raising experiment reported in Onderstal, Schram and Soetevent (2013), the authors themselves took to the streets to distribute the flyer announcing the fundraising drive. Key between-treatment differences existed in the informational content of the flyer. Households in close geographical proximity (opposite sides of the street) were assigned to different treatments. Therefore, the researchers feared that student assistants would easily make mistakes (putting the wrong flyer in the mailbox) without noticing themselves or without notifying the authors for reasons of embarrassment.

In the analysis stage, one can deal with non-compliance by conducting an intention-to-treat (ITT) analysis that compares the groups created by the initial randomization and ignores non-compliance (Mealli and Rubin, 2002). ITT is a measure of the average effect of assignment to the treatment group, regardless of the prevalence of non-compliance. ITT estimates may be conservative if many participants did not receive the assigned treatment for reasons related to the experiment, and not to the program being evaluated. Also, depending on the specific research question, it might be quite important to measure the effect of the treatment instead of the intention to treat. If certain conditions are satisfied, one can use initial assignment as an instrumental variable to estimate the effect of the treatment on a specific group of individuals. See Angrist and Imbens (1994, 1995) for a detailed discussion on this issue.

Another, related, phenomenon is attrition. Attrition occurs when outcome data are missing, for example because participants drop out of the experiment or when third-party collaborators (government agencies or firms) block access to data. Attrition necessarily reduces the sample size and thereby the power of the design. In addition, if attrition is non-random, the difference in average outcome in the treatment and control group becomes a biased estimator of the treatment effect. Unfortunately, the assumption that outcome data are missing at random is non-refutable for the simple reason that the available data reveal nothing about the distribution of the missing data (Manski, 2007, Section 2.5). This makes attrition a difficult problem to solve ex post.

Attrition is for obvious reasons more of an issue in field experiments than in laboratory experiments where strict rules on participation can be imposed and subjects have a monetary incentive to complete the experiment. How can attrition in the field be mitigated in the design stage? First and foremost, researchers should make efforts to maintain the interest of participants in participating. For example, the advantage of a phase-in design is that participants in the control group know that eventually they will also receive treatment. This may motivate them not to drop out. Outside partners should maintain their interest throughout such that they do not suddenly reduce the time its project manager or data manager can devote to the project. To accomplish this, researchers could provide regular progress reports on the project. In the analysis stage, one can address attrition by first determining the potential size of attrition and using this information to derive lower and upper bounds on the size of the treatment effect.

## 6 FINAL REMARKS

In this chapter, we discussed several important topics related to randomization in field experiments. In Section 2, we reviewed how randomization can help estimating treatment effects. In Section 3.1, we discussed how spillovers in the field can challenge the validity of SUTVA. We looked at several articles where different levels of randomization were exploited to address spillovers, and we have also explored other solutions. In Section 3.2, we focused on implementation and practical issues that matter in selecting the level of randomization. In Section 3.3, we discussed power calculations and highlighted how a higher level of randomization reduces power.

In Section 4, we turned our attention to the use of covariates. We explained why stratification is advantageous and gave examples of stratified designs from the field. We

also touched upon the use of covariates in the analysis stage. In Section 5, we discussed two ways in which plans could fail. The first was that the assignment mechanism might not be fully controlled, and the second was that treatment compliance is only partial. In both cases, we looked for ways in which the problem can be prevented, and discussed some possible solutions that can be applied if the problem did occur.

Of course, we could not cover many other interesting topics due to space constraints; for example, the number of treatment arms or cross-cutting designs. There is also more to say about some of the topics that we did include, but a complete treatment of all aspects is outside the scope of our chapter. We hope that our overview has invigorated the reader, and we close by providing references for further reading for those who are thirsty for more. Duflo et al. (2007) provide a toolkit to researchers who would like to use randomized experiments in development economics. Gerber and Green (2008) is a very accessible treatise on the strengths and weaknesses of field experimentation that discusses applications in political science. Gandhi et al. (2016) provide guidelines for researchers who wish to run field experiments to evaluate energy efficiency programs. Baird et al. (2018) give an in-depth and formal discussion on experiments with a two-step design. Finally, Duflo and Banerjee (2017) is a recent two-volume handbook dedicated solely to field experiments.

## REFERENCES


Allcott, Hunt and Todd Rogers. 2014. "The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation." *American Economic Review* **104**(10): 3003–37.

Angrist, Joshua D. and Guido W. Imbens. 1994. "Identification and estimation of local average treatment effects." *Econometrica* **62**(2): 467–75.

Angrist, Joshua D. and Guido W. Imbens. 1995. "Two-stage least squares estimation of average causal effects in models with variable treatment intensity." *Journal of the American Statistical Association* **90**(430): 431–42.

Ashraf, Nava, Erica Field and Jean Lee. 2014. "Household bargaining and excess fertility: an experimental study in Zambia." *American Economic Review* **104**(7): 2210–37.

Athey, Susan and Guido W. Imbens. 2017. "The econometrics of randomized experiments." In Esther Duflo and Abhijit Banerjee (eds), *Handbook of Field Experiments, Vol. 1* (pp. 73–140). Amsterdam: North-Holland.

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh and Berk Ozler. 2018. "Optimal design of experiments in the presence of interference." *Review of Economics and Statistics* **100**(5): 844–60.

Bhargava, Saurabh and Danyanand Manoli. 2015. "Psychological frictions and the incomplete take-up of social benefits: evidence from an IRS field experiment." *American Economic Review* **105**(11): 3489–529.

Bloom, Howard S. 1995. "Minimum detectable effects: a simple way to report the statistical power of experimental designs." *Evaluation Review* **19**: 547–56.

Bloom, Howard S. (ed.). 2005. "Learning more from social experiments." In *Randomizing Groups to Evaluate Place Based Programs* (pp. 115–72). New York: Russell Sage Foundation.

Bloom, Nicholas, James Liang, John Roberts and Zhichun Jenny Ying. 2015. "Does working from home work? Evidence from a Chinese experiment." *Quarterly Journal of Economics* **130**(1): 165–218.

Booij, Adam S., Edwin Leuven and Hessel Oosterbeek. 2017. "Ability peer effects in university: evidence from a randomized experiment." *Review of Economic Studies* **84**(2): 547–78.

Bursztyn, Leonardo, Florian Ederer, Bruno Ferman and Noam Yuchtman. 2014. "Understanding mechanisms underlying peer effects: evidence from a field experiment on financial decisions." *Econometrica* **82**(4): 1273–301.

Callen, Michael and James D. Long. 2015. "Institutional corruption and election fraud: evidence from a field experiment in Afghanistan." *American Economic Review* **105**(1): 354–81.

Carneiro, Pedro, Sokbae Lee and Daniel Wilhelm. 2017. "Optimal data collection for randomized control trials." *CEMMAP Working Paper CWP15/17*.

Crépon, Bruno, Esther Duflo and Marc Gurgand et al. 2013. "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment." *Quarterly Journal of Economics* **128**(2): 531–80.

DellaVigna, Stefano, John A. List and Ulrike Malmendier. 2012. "Testing for altruism and social pressure in charitable giving." *Quarterly Journal of Economics* **127**(1): 1–56.

Duflo, Esther and Abhijit Banerjee (eds). 2017. *Handbook of Field Experiments, Vols 1 & 2*, Amsterdam: North-Holland.
Duflo, Esther, Rachel Glennerster and Michael Kremer. 2007. "Using randomization in development economics research: a toolkit." In Paul Schultz and John A. Strauss (eds), *Handbook of Development Economics, Vol. 4* (pp. 3895–962). Amsterdam: North-Holland.
Duflo, Esther, Rema Hanna and Stephen P. Ryan. 2012. "Incentives work: getting teachers to come to school." *American Economic Review* **102**(4): 1241–78.
Dupas, Pascaline. 2014. "Short-run subsidies and long-run adoption of new health products: evidence from a field experiment." *Econometrica* **82**(1): 197–228.
Dupas, Pascaline and Edward Miguel. 2017. "Impacts and determinants of health levels in low-income countries." In Esther Duflo and Abhijit Banerjee (eds), *Handbook of Field Experiments, Vol. 2* (pp. 3–93). Amsterdam: North-Holland.
Faul, Franz, Edgar Erdfelder, Albert G. Lang and Axel Büchner. 2007. "G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences." *Behavior Research Methods* **39**: 175–91.
Fosgaard, Toke R. and Adriaan R. Soetevent (2018). "Promises undone: how committed pledges impact donations to charity." *Tinbergen Institute Discussion Paper No. 18-044/VII.*
Friebel, Guido, Matthias Heinz, Miriam Krueger and Nikolay Zubanov. 2017. "Team incentives and performance: evidence from a retail chain." *American Economic Review* **107**(8): 2168–203.
Fryer, Roland G. 2014. "Injecting charter school best practices into traditional public schools: evidence from field experiments." *The Quarterly Journal of Economics* **129**(3): 1355–407.
Gandhi, Raina, Christopher R. Knittel, Paula Pedro and Catherine Wolfram. 2016 "Running randomized field experiments for energy efficiency programs." *Economics of Energy & Environmental Policy* **5**(2): 7–24.
Gerber, Alan S. and Donald P. Green. 2008. "Field experiments and natural experiments." In Janet M. Box-Steffensmeier, Henry E. Brady and David Gollier (eds), *The Oxford Handbook of Political Methodology* (pp. 357–403). Oxford: Oxford University Press.
Haushofer, Johannes and Jeremy Shapiro. 2016. "The short-term impact of unconditional cash transfers to the poor: experimental evidence from Kenya." *Quarterly Journal of Economics* **131**(4): 1973–2042.
Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* **81**(396): 945–60.
Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social and Biomedical Sciences*. New York: Cambridge University Press.
Kendall, Chad, Tommaso Nannicini and Francesco Trebbi. 2015 "How do voters respond to information? Evidence from a randomized campaign." *American Economic Review* **105**(1): 322–53.
Manski, Charles F. 2007. *Identification for Prediction and Decision*. Boston, MA: Harvard University Press.
Mealli, Fabrizia and Donald B. Rubin. 2002. "Assumptions when analyzing randomized experiments with noncompliance and missing outcomes." *Health Services & Outcomes Research Methodology* **3**: 225–32.
Muralidharan, Karthik and Venkatesh Sundararaman. 2015. "The aggregate effect of school choice: evidence from a two-stage experiment in India." *Quarterly Journal of Economics* **130**(3): 1011–66.
Onderstal, Sander, Arthur J.H.C. Schram and Adriaan R. Soetevent. 2013. "Bidding to give in the field." *Journal of Public Economics* **105**: 72–85.
Raudenbusch, Stephen, Jessica Spybrook and Howard Bloom et al. 2011. *Optimal Design Software for Multilevel and Longitudinal Research (Version 3.01)* [software]. Available from www.wtgrantfoundation.org.
Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* **66**(5): 688–701.
Voors, Maarten J., Eleonora E.M. Nillesen and Philip Verwimp et al. 2012. "Violent conflict and behavior: a field experiment in Burundi." *American Economic Review* **102**(2): 941–64.