CHAPTER 5

The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency

R. Glennerster
Massachusetts Institute of Technology, J-PAL, Cambridge, MA, United States
E-mail: rglenner@mit.edu

Contents

Handbook of Economic Field Experiments, Volume 1
ISSN 2214-658X, http://dx.doi.org/10.1016/bs.hefe.2016.10.002

Abstract

A number of critical innovations spurred the rapid expansion in the use of field experiments by academics. Some of these were econometric but many were intensely practical. Researchers learned how to work with a wide range of implementing organizations from small, local nongovernmental organizations to large government bureaucracies. They improved data collection techniques and switched to digital data collection. As researchers got more involved in the design and implementation of the interventions they tested, new ethical issues arose. Finally, the dramatic rise in the use of experiments increased the benefits associated with research transparency. This chapter records some of these practical innovations. It focuses on how to select and effectively work with the organization running an intervention which is being evaluated; ways to minimize attrition, monitor enumerators, and ensure data are collected consistently in treatment and comparison areas; practical ethical issues such as when to start the ethics approval process; and research transparency, including how to prevent publication bias and data mining and the role of experimental registries, preanalysis plans, data publication reanalysis, and replication efforts.

Keywords

Data collection; Ethics; Field experiments; Partnerships; Research transparency

JEL Codes

C81; C93; O10; O12; O22

Economists have known for a long time that randomization could help identify causal connections by solving the problem of selection bias. In chapter "The Politics and Practice of Social Experiments: Seeds of a Revolution" by Gueron (2017) and Gueron and Rolston (2013) describe the effort in the United States to move experiments out of the laboratory into the policy world in the 1960s and 1970s. This experience was critical in proving the feasibility of field experiments, working through some of the important ethical questions involved, showing how researchers and practitioners could work together, and demonstrating that the results of field experiments were often very different from those generated by observational studies. Interestingly, there was relatively limited academic support for this first wave of field experiments (Gueron and Rolston, 2013), most of which were carried out by research groups such as MDRC, Abt, and Mathematica, to evaluate US government programs, and they primarily used individual-level randomization. In contrast, a more recent wave of field experiments started in the mid-1990s was driven by academics, initially focused on developing countries, often worked with nongovernmental organizations, and frequently used clustered designs.

A number of critical innovations spurred the take-off of field experiments, particularly in academic circles. Some of these were theoretical: They included understanding how to maximize power from limited sample sizes (Imbens, 2011; Bruhn and McKenzie, 2009); how to use randomized control trials (RCTs) to measure externalities (Miguel and Kremer, 2004); the diffusion of information (Duflo and Saez, 2002; Kremer and Miguel, 2007); equilibrium effects (Crépon et al., 2012; Mobarak and Rosenzweig, 2014); and parameters in network theory (Chandrasekhar et al., 2015; Beaman et al., 2013).

Many of the innovations that powered the growth of field experiments, however, were intensely practical. Researchers learned how to work with a wide range of implementing organizations including local nongovernmental organizations, private companies, and social entrepreneurs. Unlike governments, with whom most early RCTs were conducted, these new partners tended to be more open to trying new approaches to solving problems and were more willing to test different aspects of their programs separately and in combination. Logistical and financial constraints meant they could not reach everyone they wanted to, making randomization a natural method for allocating rationed resources. There were also important practical innovations in measurement which opened up new subject areas to field experiments. With these new partners and new subject areas for experiments came a range of new ethical questions, including how to define the boundary between practice and research and how to regulate activity across that boundary as researchers got more and more involved in the design and implementation of the interventions they tested. Finally, the dramatic rise in the use of experiments increased the benefits associated with research transparency.

This chapter seeks to record some of these practical innovations that have accompanied and enabled the expansion in the use of field experiments. It is impossible to be comprehensive, so we focus on four discrete and important issues. Section 1 discusses how to select and work with a partner organization that will implement the program to be evaluated by an RCT, as well the conditions under which it makes sense for a researcher to be both the implementer and the evaluator of a program. Section 2 discusses practical challenges in data collection and strategies to combat them, including minimizing attrition, monitoring enumerators, and ensuring that data are collected consistently in treatment and comparison areas. Section 3 covers the practical ethical issues a researcher conducting randomized evaluations must take into account when designing and carrying out their research. Section 4 covers topics in research transparency, including publication bias, data mining, experimental registries, preanalysis plans (PAPs), data publication reanalysis, and replication.

1. COLLABORATION BETWEEN RESEARCHERS AND IMPLEMENTERS

Unlike most academic economic research, running field-based RCTs often involves intense collaboration between researchers and the organization or individuals who are implementing the intervention that is being evaluated. This collaboration can be the best thing about working on a field experiment or the worst. If the collaboration goes well, the researcher can learn an enormous amount from the implementing partner about how local formal and informal institutions work, how to measure outcomes in the local context, and how to interpret the results of the study. If the partnership is going badly it is almost impossible to run a high-quality field experiment. In this section we discuss practical ways to develop and maintain a good collaboration with an implementing partner.

We start with tips on how to find the right implementer and what to do to make the researcher–implementer partnership as effective as possible. We then examine whether and when it is worth attempting to "self-implement," i.e., be both implementer and evaluator.

1.1 Developing a good researcher–implementer partnership

Researcher and implementer partnerships, like any other relationship, require listening to and understanding the other partner, being flexible to their needs, respecting the other's contribution, and being honest. During an initial "courtship" phase, the two groups seek to understand whether they want to enter into an evaluation partnership. What should a researcher be looking for in an implementer during this phase? What can a researcher do to make themselves useful to, and thus support a good relationship with, an implementing organization?

1.2 What makes a good implementing partner?

1. **Sufficient scale**

   A first and easy-to-determine filter for a good implementing partner is whether an organization is working at a big enough scale to be able to generate a sample size that will provide enough power for the experiment. How big is sufficient depends on the level at which the randomization is going to take place (see chapter: The Econometrics of Randomized Experiments by Athey and Imbens (2017), as well as the number of variants of the program that are going to be compared, and the outcome of interest. Thus a lot of detailed discussion needs to take place about what a potential evaluation would look like before it is possible to say if an evaluation is feasible. It is surprising how many potential partnerships can be ruled out quite early on because the implementer is just not working at a big enough scale to make a decent evaluation possible.

2. **Flexibility**

   A willingness to try different versions of the program and adapt elements in response to discussions with researchers is an important attribute of an implementing partner. We can learn a lot by testing different parts of a program together and separately or by comparing different approaches to the same problem against each other, but doing this type of testing requires a very flexible partner. The best partnerships are the ones in which researcher and implementer work together to decide the most interesting versions of the program to test.

3. **Technical programmatic expertise and a representative program**

   There is a risk of testing a program run by an inexperienced implementer, finding a null result, and generating the response, "Of course there was no impact, you worked with an inexperienced implementer." The researcher also has less to learn about how good programs are run from an inexperienced implementer and the

partnership risks becoming one sided. At the other end of the spectrum, we may not want to work with a gold-plated implementer unless we want to test proof of concept. There are two risks here: that the program is so expensive that it will never be cost-effective even if it is effective; and that it relies on unusual and difficult-to-reproduce resources. For example, a program that relies on a few very dynamic teachers or mentors might be hard to replicate. An implementer working at a very big scale is unlikely to run a gold-plated program and has already shown the program can be scaled. It is also possible to work with a smaller implementer, but one that closely follows a model used by others. The microcredit organization Spandana was a perfect implementation partner for our evaluation of the impact of microcredit (Banerjee et al., 2015a). They operated at a large scale and their credit product was close to that of many other microcredit organizations. We tested their impact as they expanded into a large Indian city, a popular type of location for microcredit organizations.

4. **Local expertise and reputation**

Implementers who have been working with a population for many years have in-depth knowledge of local formal and informal institutions, population characteristics, and geography that is invaluable in designing and implementing an evaluation. They can answer questions, such as what messages are likely to resonate with this population? what does success look like and how can we measure it? When I started working in Sierra Leone I spent a long time traveling round the country with staff from Statistics Sierra Leone, Care, and the Institutional Reform and Capacity Building Project. One had worked with Paul Richards, an important anthropologist in Sierra Leone. Our final measures of trust, group membership, and collective action relied heavily on their suggestions and input. I learned that it was socially acceptable to ask about the bloody civil war that had just ended but that asking about marital disputes could get us thrown out of the village. From Tejan Rogers I learned that every rural (and some urban) communities in Sierra Leone come together for "road brushing" where they clear encroaching vegetation from the dirt road that links their community to the next and even build the common palm-log bridges over rivers. How often this activity took place and the proportion of the community that took part became our preferred measure of collective action and has been used by many other authors since.

Just as importantly, an implementer who has been working locally has a reputation in local communities that would take a researcher years to build. This reputation can be vital. We learn little about the impact of a program if suspicion of the implementer means that few take up the program. The reputation of the implementing organization can also be critical to the research team being permitted to operate in the community and to getting a high response rate to surveys.

Researchers need to understand how valuable this reputational capital is to the implementer. What may seem like reluctance on the part of the implementing partner to try new ideas may be a fully justified caution to put their hard-won reputation on the line.

5. **Low staff turnover**

Evaluation is a partnership of trust and understanding and this takes time to build. All too often a key counterpart in the implementing organization will move on just as an evaluation is reaching a critical stage. Their successor may be less open to evaluation, want to test a different question, be against randomization, or just uninterested. High turnover can happen in any organization, but governments and organizations with foreign staff are particularly likely to have high turnover. NGOs that draw their staff from the local community tend to experience less staff turnover. The only way a researcher can protect the evaluation is to try and build relationships at many levels of the implementing organization, so that the loss of one champion does not doom the entire project.

6. **Desire to know the truth and willingness to invest in uncovering it**

The most important quality of an implementing partner is the desire to know the true impact of an intervention and a willingness to devote time and energy to helping the researcher uncover the truth. Many organizations start off enthusiastic about the idea of an evaluation: they want an expert to certify that their program is very successful. At some point these organizations realize that it is possible that a rigorous evaluation may conclude that their program does not have a positive impact. At this point, two reactions are possible: a sudden realization of all the practical constraints that will make an evaluation impossible or a renewed commitment to learn.

In Glennerster and Takaravasha (2013, p. 20), we quote Rukmini Banerji of Pratham at the launch of an evaluation of Pratham's flagship "Read India" program:

> *And of course [the researchers] may find that it doesn't work. But if it doesn't work, we need to know that. We owe it to ourselves and the communities we work with not to waste their and our time and resources on a program that does not help children learn. If we find that this program isn't working, we will go and develop something that will.*[1]

This is the kind of commitment that makes an ideal partner. It is not just that an unwilling partner can throw obstacles in the path of an effective evaluation. An implementation partner needs to be an active and committed member of the evaluation team. There will inevitably be problems that come up during the evaluation process that the implementer will have to help solve, often at a financial or time cost to themselves. The baseline may run behind schedule and implementation will need to be delayed until it is complete; transport costs of the program might be higher

[1] This quote reflects my memory of Rukmini's speech.

as implementation communities end up being further apart than they otherwise would be to allow for controls; when and where a program is to be rolled out may need to be set further in advance because of the evaluation; selection criteria must be written down and followed scrupulously to reduce the discretion of local staff in accepting people into the program; or some promising program areas may need to be left for the control group. Partners will only put up with these problems and actively help solve them if they fully appreciate the benefits of the evaluation being high quality and if they understand why these restrictions are necessary to a high-quality evaluation. Padmaja Reddy of Spandana provides a good example of this commitment. In the early stages of our evaluation of Spandana's microcredit product we became aware that credit officers from Spandana were going into some control areas to recruit microcredit clients. Only Padmaja's active intervention managed to stop this activity, which would have undermined the entire experiment if left unchecked.

Commitment to the evaluation needs to come from many levels of the organization. If the headquarters in Delhi want to do an impact evaluation but the local staff do not, it is not advisable for HQ to force the evaluation through because it is the staff at the local level who will need to be deeply involved in working through the details with the researcher. Similarly, if the local staff are committed but the HQ is not, there will not be support for the extra time and cost the local staff will need to expend to participate in the study. Worst of all is when a funder forces an unwilling implementer to do an RCT run by a researcher. My own and others' bitter experience suggests that being involved in a scenario of this kind will suck up months of a researcher's time trying to come up with evaluation designs that the implementer will find some way to object to.

If this level of commitment to discovering the unvarnished truth sounds a little optimistic, there are practical ways to make an impact evaluation less threatening to a partner. An implementer who runs many types of programs has less at stake from an impact evaluation of one of their programs than an organization with a single signature program. Another option is to test different variants of a program rather than the impact of the program itself. For example, testing the pros and cons of weekly versus monthly repayment of microcredit loans (Field et al., 2012) is less threatening than testing the impact of microcredit loans. In some cases researchers have started relationships with implementers by testing a question that is less threatening (although potentially less interesting). As the partnership has built up trust, the implementing partner has opened up more and more of their portfolio to rigorous testing.

7. Trade-offs between partner criteria

An important concern is that a partner, who is committed to knowing the truth about the effect of the program understands randomization, and has the time and expertise to invest in a serious evaluation partnership, is unlikely to be representative

of other implementers. We may worry that there is a systematic bias in the programs that are evaluated. Allcott (2015) examines 111 RCTs of similar programs to encourage energy conservation across the United States. He finds that the program was more effective in the first 10 sites to adopt the program and be evaluated than those who adopted and were evaluated later. This holds true even after correcting for observable differences between sites. He suggests that utilities in areas that were particularly keen to reduce energy signed up to the program earlier and also had clients who responded more to conservation messages. Note that in this case evaluation and program adoption are a single package. Allcott's estimation of site selection bias combines two possible biases: the program is more effective for those who (1) are early adopters and (2) are willing to be evaluated. Allcott is not able to test whether those who are willing to evaluate are likely to run higher quality programs because, in his case, the programs are all run by a single operator at different sites.

Whether we want to prioritize having a representative partner or a highly committed partner depends on the objective of the research. If we are testing an underlying human behavior—such as a willingness to pay now for benefits in the future—the representativeness of the partner may be less relevant. If we want to know whether a type of program, as it is usually implemented, works, we will want to prioritize working with a representative partner. Note that "does this type of program work" is not necessarily a more policy-relevant question than a more general question about human behavior. By their nature, more general questions generalize better and can be applied to a wider range of policy questions.

1.3 What can a researcher do to foster a good partnership with an implementing organization?

We have set out a long list of characteristics a researcher wants in an implementing partner. But what does an implementer want in a research partner, and how can a researcher make him- or herself a better partner?

1. **Answer questions the partner wants to be answered**

   Start by listening. A researcher will go into a partnership with ideas about what they want to test, but it is important to understand what the implementer wants to learn from the partnership. Try to include a component of the evaluation that answers the key questions of the implementer as well as elements that answer the key researcher questions. For example, sometimes these questions do not require another arm to be added to the study, but rather some good monitoring data or quantitative descriptive data of conditions in the population to be collected.

2. **Be flexible about the evaluation design**

   The research design a researcher has in his/her head when he/she starts a partnership dialogue is almost never the design that ends up being implemented. It is critical to respond flexibly to the practical concerns raised by the implementer. One of the

main reasons that randomized evaluations have taken off in development in the last 20 years is because a range of tools have been developed to introduce an element of randomization in different ways. It is important to go into a conversation with a partner with all those tools in mind and use the flexibility they provide to achieve a rigorous study that also takes into account the concerns of the implementer.

A common concern implementers have about randomization is that they will lose the ability to choose the individuals or communities that they think are most likely to benefit from their intervention. They may worry a community mobilization program will not work if the community is too large and lacks cohesiveness, or is too small to have the resources to participate fully. A training program may want to enroll students who have some education but not too much. These concerns are relatively easy to deal with: agree to drop individuals or communities that do not fit the criteria as long as there are enough remaining to randomize some into treatment and some into control. This may require expanding the geographic scope of the program. Randomization in the bubble can be a useful design in dealing with these concerns.

Randomized phase-in designs are also useful for addressing implementer concerns, although they come with important downsides (Glennerster and Takavarasha (2013) detail the pros and cons of different randomization techniques.).

There are limits to the flexibility that can and should be shown. If an implementing organization repeatedly turns down many different research designs that are carefully tailored to address concerns that have been raised in previous conversations, at some point the researcher needs to assess whether the implementer wants the evaluation to succeed. This is a very hard judgment to make and is often clouded by an unwillingness to walk away from an idea that the researcher has invested a lot of time in. The key question to focus on in this situation is whether the implementer is also trying to overcome the practical obstacles to the evaluation. If not, then it probably makes sense to walk away and let go of the sunk costs already invested. Better to walk now than be forced to walk away later when even more time and money have been invested.

3. **Share expertise**

Many partners are interested in learning more about impact evaluation as part of the process of engaging with a researcher on an evaluation. Take the time to explain the impact of evaluation techniques to them and involve them in every step of the process. Offer to do a training on randomized evaluations for staff at the organization or run a workshop on Stata. Having an organization-wide understanding of randomized evaluations also has important benefits for the research. In Bangladesh, employees of the Bangladesh Development Society were so well versed in the logic of RCTs that they intervened when they noticed girls attending program activities from surrounding communities. They explained to the communities (unprompted) that this could contaminate the control group and asked that only local girls attend.

Researchers often have considerable expertise in specific elements of program design, including monitoring systems and incentives, as well as knowing about potential sources of funding—all of which can be highly valued by implementers. Many researchers end up providing technical assistance on monitoring systems and program design that go well beyond the program being evaluated. The goodwill earned is invaluable when difficult issues arise later in the evaluation process.

4. **Provide intermediate products**

   While implementing partners benefit from the final evaluation results, the timescales of project funding and reporting are very different from academic timelines. Often an implementing organization will need to seek funding to keep the program going before the end line is in place and several years before the final evaluation report is complete. It is therefore very helpful to provide intermediate outputs. These can include a write-up of a needs assessment in which the researcher draws on existing data and/or qualitative work that is used in project design; a description of similar programs elsewhere; a baseline report that provides detailed descriptive data of the conditions at the start of the program; or regular monitoring reports from any ongoing monitoring of project implementation the researchers are doing. Usually researchers collect these data but do not write them up until the final paper. Being conscious of the implementers' different timescale and getting these products out early can make them much more useful.

5. **Have a local presence and keep in frequent contact**

   Partnerships take work and face time. A field experiment is not something you set up, walk away from, and come back sometime later to discover the results. Things will happen, especially in developing countries: strikes, funding cuts, price rises, Ebola outbreaks. It is important to have a member of the research team on the ground to help the implementing partner think through how to deal with minor and major shocks in a way that fits the needs of both the implementer and the researcher. Even in the middle of multiyear projects I have weekly calls with my research assistants, who either sit in the offices of the implementer or visit them frequently. We always have plenty to talk about. I also visit the research site once and often twice a year. Common issues that come up during the evaluation are lower-than-expected program take-up, higher-than-expected costs of running the program, uneven implementation quality, and new ideas on how to improve the program.

1.4 Special considerations when partnering with governments

Working with government partners has particular benefits and challenges. On the benefit side, governments often have substantial resources at their disposal and their geographic reach is expansive. Thus, for example, Olken et al. (2014) were able to randomize at the level of subdistrict in Indonesia with 1.8 million target beneficiaries in treatment areas.

Governments also collect a lot of data on individuals such as test scores for children, earnings for adults, and encounters with the criminal justice system. While it may be possible to access these data even if the government is not the implementer of the program being evaluated, a formal partnership makes doing so much easier. Administrative data can enable researchers to assess impacts without extensive surveys.

This is particularly beneficial for study designs that require large samples sizes and/or long-term tracking. For example, Angrist et al. (2006) are able to follow up with winners and losers of a lottery for vouchers to attend private school in Colombia by linking winners to a centralized college entry exam seven years after the vouchers were issued. In ongoing work, Bettinger et al. link the same voucher winners and losers to government tax and earnings data, 17 years after the lottery. Governments' wide reach makes it possible to randomize on populations that are representative of large geographic units. Muralidharan and Sundararaman (2011) test the impact of teacher incentive pay in a representative sample of rural schools across the state of Andhra Pradesh, meaning their results are valid across a population of 60 million.[2]

Another benefit of working with governments is that they have the ability to scale-up a program to a large number of people if a pilot is found to be effective. If the evaluation is of a government-implemented pilot this may ease, though not necessarily erase, the concern that the scale-up will not be implemented as well as the pilot. Governments may also find the results from such a pilot more persuasive than one conducted by another organization. In 2015, Banerjee, Hanna, and Olken worked with the Government of Indonesia to test how providing individual ID cards to recipients of government-subsidized rice (which indicated the amount and price of rice they were eligible for) could reduce corruption in the distribution system. The results showed that the cards increased the subsidy received by targeted recipients by 25%, so the government scaled up the ID card program, reaching 66 million people. The time from evaluation design to scale-up was about a year.

Some issues can only be examined by working with governments: for example, manipulating how tax collectors are rewarded (Khan et al., 2014); how police are trained and rewarded (Banerjee et al., 2012); or how firms' emissions into the environment are regulated (Duflo et al., 2013).

With these benefits, however, come considerable costs. Governments can be slow moving and less able or willing to test out-of-the-box solutions than NGOs. It may be particularly difficult to run more theory-oriented field experiments with governments. They tend to be less interested in answering an abstract question, the answer to which could inform many policies but would not be scaled up as a specific program. Governments can also find it harder than NGOs to provide services only to a limited

[2] Andhra Pradesh has a population of 80 million and is 75% rural.

group of needy citizens. Some governments have laws requiring them to treat citizens of equivalent need equally. When the Government of France wanted to test programs using randomized trials they first had to change the constitution to make this possible (J-PAL, 2015). Additionally, staff turnover in governments can be high as civil servants are transferred regularly. This makes it even more important to build support at different levels of government: if the RCT has support from the minister but not the bureaucrats, then it is likely to die with the next cabinet reshuffle. An election can lead to a dramatic change in policy priorities and personnel at the same time. It can also lead to paralysis for a period both before and after an election, even if the program being evaluated has bipartisan support. As an example, an RCT I was involved in collapsed when none of the planned monitoring could take place because a newly elected government froze all nonessential expenditure while they thought through new priorities. In another instance a survey had to be suspended just as it was about to go into the field because of a national exchange rate crunch, which again led to a freeze. Government budget shortfalls and last minute crunches are not confined to developing countries. Finally, it is worth recognizing that governments can renege on any agreement with impunity. There is not much a researcher can do when a government decides to fill a shortfall in a program budget with money set aside for, say, the end line.

Many of the strategies discussed earlier for fostering partnerships in general are particularly important for fostering partnerships with governments. Government partners are in a powerful position vis-à-vis the researcher, so it is important to listen hard to what they want. They often work within short political timelines, so delivering intermediate products such as baseline reports can be key for keeping them engaged.

Working with governments often requires a more formal approach to partnership than working with NGOs. Governments often require a memorandum of understanding that sets out clear expectations for both parties. Discussions may be going well at the practical implementation level, but any final decision—even a relatively small one—is likely to require sign-off from someone senior. It is important to build extra time into the schedule to account for this. Government procurement rules can also cause considerable delay. For example, if we decide that an intervention needs a leaflet to explain the study to participants, the government may require a competitive bid for the printing of the leaflet, leading to several months delay. Having some independent funding that does not run through the government can be very helpful in easing some of these constraints: a researcher can come in and offer to pay for a leaflet, or for additional monitoring, etc. Independent funding can also help keep the research going if the government faces short-run liquidity constraints.

Being the first to do something might be exciting for an NGO but can make a government nervous as it exposes them to criticism. Thinking through the optics of the experiment (i.e., how it would look on the front page of a newspaper) can help alleviate concern. Another strategy is to bring in an official from another department or country

who has worked on an experiment before, preferably of a similar type. It is much more reassuring for officials to talk to other officials than it is to hear from a researcher.

Policymakers often have a healthy skepticism of researchers who want to provide advice about how to measure or improve a program, especially those coming from another country, state, or region. It is important for researchers to prove their relevance and their local knowledge. A mix of humility, a desire to learn from the policymaker, and a lot of homework about local conditions can help. I have seen policymakers visibly relax and start to engage when they hear from a researcher about their on-the-ground experience. A well-placed anecdote about a conversation with a farmer in Kenema or a teacher in Pittsburgh can be critical for building credibility.

1.5 Self-implementation

The major benefit of not working with an implementing partner, but implementing the intervention as a researcher, is the high degree of flexibility to precisely test the intervention or range of interventions. To understand how and why a particular program has the impact it has, we may want to take it apart and test different elements separately and together, and it may be hard to find an implementer who is willing to do this. For example, community-driven development (CDD) is a very common development program that combines the provision of block grants for locally designed projects to communities with facilitation to encourage inclusive decision making in selecting the programs. For many years, researchers have wanted to test the marginal benefit of the facilitation, but this would involve providing some grants without facilitation, something most implementers of CDD are strongly opposed to. The result is that most studies have tested the combination of grants and facilitation (Casey et al., 2012; Fearon et al., 2009; Humphreys et al., 2012; Beath et al., 2013).

We may want to compare two very different types of programs that are designed to deliver the same outcome against each other. But individual implementers may specialize in doing program A or program B, with none willing and able to do A in some randomly determined locations and B in others. We could try and find two implementers who would cooperate on where they did their respective intervention, but this kind of tripartite collaboration is likely to be exceptionally difficult. Even if we succeed it will be impossible to disentangle the differential impact of program A versus B from the impact of differential implementation skill of the organizations running each. A good example of this is the potential comparison between any program and cash. It is often useful to compare the effectiveness of a program in achieving a given objective to providing cash in achieving the same outcome. As with the CDD example mentioned in previous paragraphs, most implementers are reluctant to simply hand out cash (an exception is GiveDirectly, which was started by academic economists with the ultimately correct view that giving out cash might be an effective way to help the poor with few downsides

(Haushofer and Shapiro, 2013)). It is sometimes possible to reach a compromise with partners to do this type of comparison. A study in Bangladesh randomized different elements of Save the Children's girls' empowerment program but also added an arm with an (noncash) incentive to delay marriage. While not part of the original program, Save the Children agreed to a hybrid arrangement where the researchers took the responsibility for designing, raising the funding for, and helping to implement the noncash delivery program, while Save the Children supported the delivery of the noncash incentive through its existing food distribution system and provided support in implementation so that this element closely resembled a Save the Children program (Buchmann et al., 2016).[3]

The flexibility of self-implementation is particularly useful when we want to test a theory of underlying human behavior through an intervention that may not have a lot of practical benefit in itself. Lab experiments are an extreme form of this. Implementing partners are unlikely to want to run a lab experiment and they do not have as much expertise to contribute as this is far removed from what they normally do. But lab experiments can be very useful in testing precise hypotheses because they isolate very specific differences between arms. Many RCTs outside the lab are effectively somewhere between lab experiments and program evaluations.

A series of RCTs on take-up of health prevention products are a good example of the continuum between program evaluation and lab experiments, and how researchers shift from working with implementers to implementing themselves through this continuum. Kremer and Miguel (2007) worked with a nongovernmental organization to randomize the price at which deworming pills were provided as part of a larger program. Ashraf et al. (2010) sought to understand whether price influenced use of health products (something that was not an issue for deworming pills) and to distinguish between a psychological commitment effect of paying for a product and a selection effect. To do this, people went door to door selling dilute chlorine at randomly selected prices. Some of those who agreed to buy the chlorine at a given price then received a discount, or were surprised to receive the chlorine for free. Even though this two-stage pricing did not much resemble a normal NGO program, the researchers were able to work with Population Services International (PSI) to implement it because of the long-run relationship between the researchers and PSI and PSI's realization of the value of understanding the underlying behavior of health consumers in designing their future programs.[4] Hoffmann et al. (2009) in contrast, implemented their own program in which they randomized the price at which people were offered bed nets. To abstract from cash constraints, they

[3] A description of this ongoing study can be found at http://www.povertyactionlab.org/evaluation/empowering-girls-rural-bangladesh.

[4] Cohen and Dupas (2010) used a similar design as Ashraf et al. but with bed nets in Kenya and implemented through the research organization Innovations for Poverty Action.

provided subjects with enough cash to purchase a net prior to the offer of sale. They also looked at loss aversion by offering to purchase nets from individuals once they had bought them. While this design was very helpful in distinguishing different theories of consumer behavior with respect to preventive health, no one would think it was a good way to run a bed-net distribution program, so working with an implementing partner was unlikely to be an option (Hoffman was also a graduate student at the time, meaning she had not developed the long-run partnerships with implementing organizations that Kremer, Miguel, and Ashraf had each developed with their respective partner organizations).

Researchers sometimes choose to work through research organizations as implementers or create new implementing organizations because their empirical and theoretical works suggest a new strategy that has the potential to be effective at scale. In these cases, researchers often work through the design of the program and the research simultaneously. Chlorine Dispensers for Safe Water and StickK are examples where researchers helped create new products and organizations to scale these products, which were also evaluated through field experiments.[5]

Offsetting these important benefits of self-implementation are important disadvantages: it takes an extraordinary amount of focused attention and work to implement a complex program well; the researcher does not benefit from the insights of the implementer who usually knows a lot about the local context; questions may be raised about the extent to which the results will generalize to a program implemented not by nonresearch organizations; and different and more complicated ethical questions arise with researcher-implemented programs. As part of nonprofit universities, academics may be restricted from political advocacy, which may limit their ability to self-implement election work. (I address these last two points in Section 3.)

It is easy for researchers to underestimate the challenges in implementing a program directly, particularly in a developing country. It is common for researchers, particularly junior ones, to look at the overhead costs that implementing organizations charge and decide it would be cheaper to implement the program themselves, only to realize halfway through the experiment why others charge high overheads. Permits are hard to get, supplies do not arrive on time, staff get sick or quit, or hurricanes happen. It is hard enough to run the RCT: running the implementation at the same time is a major headache. Nor do researchers necessarily have a comparative advantage in most implementation tasks such as logistics and human-resource management. This is another reason why it is more common for researchers to self-implement the type of RCTs that have quick turnaround, and/or

[5] In the case of chlorine dispensers, the program was originally implemented by ICS Africa, then by Innovations for Poverty Action where more testing with scaling was done, before being spun off to Evidence Action. More about Dispensers for Safe Water can be found at http://www.evidenceaction.org/dispensers/. For more on StickK, see http://www.stickk.com/.

involve a lab in the field: the key tasks of implementation (such as determining the precise wording of a behavioral intervention in a lab) are closer to the comparative advantage of a researcher and long-term employment of staff is not required.

To what extent can we generalize the results from researcher-led RCTs? Vivalt (2015), in a metaanalysis of field experiments in developing countries, finds that the identity of the organization running the program is the largest predictor of impact within studies of the same type of program. This suggests that the results from a researcher-implemented program may not necessarily translate into the same impact if the program were run by a government. However, whether this is a drawback to studies of researcher-implemented programs depends a lot on what type of lesson we are seeking to draw from a study and the type of intervention that is being tested. As we have discussed, the objective of researcher-led implementation is often to tease out an underlying behavior rather than to test whether a program would be effective at scale. In this case, the fact that an NGO or government might implement the program differently than a researcher is not relevant to achieving the objectives of the study. No NGO is going to implement Hoffmann et al.'s bed net distribution the way they implemented it, but that does not undermine the general lesson about loss aversion that the RCT provides. A point that is often missed is that lessons about human behavior that often come from researcher-implemented studies or studies that are not designed to test scalable interventions are, in some ways, more generalizable than lessons from evaluations of specific programs precisely because they seek to test more general and theoretical questions.

But what if the objective of an RCT is to draw lessons about whether a particular type of intervention is effective in achieving certain outcomes and whether this type of program should be scaled? How useful are evaluations of researcher-implemented programs then? To understand this we need to think through why researcher-implemented programs may be different from those implemented by others.

Some researcher-implemented programs are criticized as not being a valid test of an approach because researchers do not have the expertise to run a program properly. One possibility is to hire someone with the technical capacity the researcher does not have. In certain disciplines (such as medicine or agronomy) an expert's qualifications can be documented and their advice can be validated by independent experts. Thus Cole and Fernando (2012) evaluated a phone-based agricultural extension program. To do this they needed an agricultural expert. The advice this expert gave is easy to assess. But in other areas this external validation of the quality of implementation is harder to do. For example, if an economics researcher ran and evaluated a program on community mobilization and found no impact, this is likely to carry less weight than a null result from a program evaluation of a well-known and respected implementer of community mobilization programs.

A more common concern is that researcher-implemented programs are not representative because they are too well-implemented. Researchers tend to have a high

level of education and, during the evaluation, will be focusing a lot of attention on a relatively small number of participants. It is, unfortunately, not typical to have so many highly educated people focus on the implementation of a program in a relatively small area. People of equivalent education levels in implementation organizations tend to be responsible for a very large number of programs often covering hundreds of thousands of people. This is not just an issue for researcher-implemented programs. Programs that are evaluated often get greater scrutiny than those that are not being evaluated. Again, however, the extent to which this is a problem depends on the objective of the study.

If a study is designed to test proof of concept, then researcher focus (as an implementer or just as an engaged partner) is not a problem. A proof-of-concept study asks the question, "What is the impact of an intervention if it is implemented as well as it could be?" Medical and public-health trials are often proof-of-concept studies. For example, it is useful to know whether addressing anemia increases productivity, even if this involves an intensive intervention in which households are given iron pills and are visited regularly to make sure there is high compliance (Thomas et al., 2003). If the study finds such a link, the question remains how best to increase iron uptake in a sustainable way. Studies of researcher-implemented programs are often proof-of-concept studies.

An alternative approach for ensuring that wider lessons can be learned from researcher-implemented studies is for the researcher to very carefully document implementation steps so that it is clear what the implementation was that was tested, and how others could replicate it. This sort of monitoring can be used to assess whether implementation quality declines as the program is scaled. This approach works best when quality is easy to measure. For example, it is possible to objectively monitor how often a chlorine dispenser is empty and therefore judge the extent to which program quality deteriorates as it is scaled and less attention is paid to each community. It is much harder to judge how the quality of a mentoring program changes as it is scaled. This point is not only relevant to researcher-implemented programs, but it is particularly relevant for them.

In summary, then, when deciding whether to implement a program as a researcher it is important to think through the objectives of the study. If it is a short lived, small-scale experiment with quite theoretical objectives where subtle differences in implementation are crucial to the design, self-implementation may be a good approach. If the objective is to test a proof of concept, and there are objective ways to measure the quality of implementation, then self-implementation may be possible—but not necessarily advisable—given the work involved. But for the vast majority of field experiments, the benefits of self-implementation do not outweigh the costs. In particular, researchers isolate themselves from potentially useful partners.

Some commentators have concluded that the involvement of researchers in the implementation of programs raises important ethical issues. The issue has arisen mainly in the context of field experiments around elections.[6] We discuss this in Section 3.

2. PREPARING FOR PRACTICAL PITFALLS IN FIELD EXPERIMENTS

When running a field experiment it is best to prepare for the worst. Some crises cannot be foreseen: in one 12-month period my field experiments were hit by Ebola, riots, a national strike, and a coup. The likelihood of unforeseen shocks makes it more critical to prepare for challenges that can be foreseen. Even with the best implementing partner, there will be issues of compliance with the randomization protocol and take-up will be lower than you expect. Even with a team of experienced and well-trained enumerators, someone will try to make up the data and attrition will have to be addressed. In this section we discuss strategies to combat these challenges.

2.1 Noncompliance

Despite our best efforts, there will always be people randomized to receive treatment who do not access the program and those who are randomized to the comparison group who manage to get access. Intention-to-treat estimates are still valid if there is a low level of noncompliance, but by reducing the contrast between treatment and comparison, noncompliance dramatically reduces power.[7] Choosing the right partner, as discussed earlier, is key to compliance but so is designing a randomization protocol that is easy to follow. Program implementers have enough to deal with in making the program run smoothly. In the best designs they have no decisions to make related to the randomization protocol. Often the best strategy is to ensure that any front-line implementer works entirely with either treatment or control people, but never implements differently across arms. For example, Buchmann et al. (2016) compared two different versions of an empowerment program, but field supervisors who supervised many villages were always given villages running the same version of the program. The one case where it is sometimes possible to have front-line staff implement differently with different people is if they are following a script on a computer and the computer randomizes the script (Duflo et al., 2005). Karlan and Appel (2016) have many examples of field experiments gone wrong, many of which involved noncompliance. There are several cases of attempts

[6] A get-out-the-vote field experiment in Montana caused considerable debate about research ethics when the fliers used in the experiment inappropriately used the Montana State seal. However, questions were also raised about whether it was ethical to conduct research that might influence the outcome of an election. For further discussion see, for example: https://thewpsa.wordpress.com/2014/10/25/messing-with-montana-get-out-the-vote-experiment-raises-ethics-questions/.

[7] The minimum detectable effect (MDE) size is squared when it enters the power equation, so power is particularly sensitive to changes in the MDE.

to evaluate layering an additional element onto an existing program. The new element was simply added to the work load of existing program staff, was not their main focus nor expertise, and as a result the new element was implemented poorly and inconsistently.

Even if program staff are implementing the program well and in line with the randomization protocol, take-up of the program may be disappointing. In Banerjee et al. (2015a), Spandana predicted that 80% of eligible women would take up their microcredit product. In planning the study we assumed this was an overestimate and predicted take-up of 60%. The actual take-up was less than 20%. Take-up is critical to power, so it is important to get right. If the program is being run elsewhere, one approach is to collect data on actual take-up in the other location. Alternatively, running a pilot on a small scale prior to the evaluation is useful for estimating take-up as well as to sort out the details of program and research implementation. Even then, take-up is likely to be lower in the main study than in the pilot as pilots often get a particularly high level of attention.

Another driver of noncompliance is that randomization units which appear separable on paper are much messier on the ground. The clear clinic catchment areas delineated on maps by the ministry of health may bear little relationship to who actually attends which clinic. Government-imposed political boundaries such as towns, villages, or states do not always correspond to the patterns of daily interaction that are likely to drive program implementation and spillovers. Even the definition of household is not always straightforward: households are usually defined as those who eat together, but it may be hard to treat one part of a family and not the other part if they live in the same dwelling, even if they do not eat together. To prevent this type of noncompliance it is critical to establish natural randomization units that are informed by how people actually interact—not how they are meant to interact according to some government plan (Chapter 4 of Glennerster and Takavarasha (2013) covers this issue in greater depth).

The most problematic form of noncompliance is the defier: people who take up treatment because they were randomized to control, or who do not take it up because they were randomized to treatment. Unlike other forms of noncompliance, which just reduce power, defiers can bias results. Defiers are most likely to occur in information interventions because of the interaction of information provision with previous priors.[8] If we are concerned about defiers we need to identify groups where they may be an issue (for example, by collecting baseline data on existing priors) and calculating heterogeneous effects. We can separately examine the effects of an information program on those where

[8] For example, we might tell people about the benefits of wearing a seat belt as a way of increasing the use of seat belts, and thus further measuring their effectiveness. However, if some people previously had an overinflated view of the benefits of seat belts, the information might actually make them less likely to use seat belts. Our estimate of the information program would be valid, but the estimate of the effect of seat belts would be incorrect if defiers are different from nondefiers in other aspects of their behavior.

the information was in line with previous priors, was higher than previous priors, or lower than priors (see Glennerster and Takavarasha (2013) for more details on this subject).

Even with the best preparation possible, things will go wrong. It is therefore essential to monitor compliance and take-up throughout the implementation phase and provide feedback to the implementer so that they can fix any issues with implementing staff and redouble efforts at take-up. Data on who is not taking up the program can be very helpful to implementers in focusing their take-up promotion strategies. Collecting data in the end line about who took up the program will be important in interpreting the results; for example, distinguishing between limited impact being driven by low take-up or by low impact among those who took up. These data are also needed for calculating treatment-on the-treated estimates where appropriate.

2.2 Attrition

A high attrition rate can ruin an otherwise well-designed and implemented RCT. Most RCTs involve collecting panel data on the same people before and after the start of the intervention. While it is possible to account for attrition in these studies by placing bounds on the estimated coefficient, unless attrition is very small, these bounds will be large, making it hard to draw precise conclusions from the results. Even RCTs that do not collect panel data still have to worry about attrition from the selected sample: if we randomly select people who were subject to a natural field experiment to measure its effects, but only reach a portion of those we sought to interview, our results could be biased.

The following are some tips for keeping attrition low:

1. **Plan for more than one visit**

   Whether the surveys are conducted in people's homes, schools, or workplaces, some people will be absent on the day the enumerators come for the survey even when they have been warned in advance. Up to three separate visits may be needed to ensure that a high proportion of people are reached.

2. **Track people where they are**

   Simply returning to the same location repeatedly may not be sufficient if the respondent has moved. If children have dropped out of school, the enumerator needs to go to the child's home, and if the outcome is child test scores, the test will need to be administered at home. If families have moved it may be necessary to track and interview people in their new locations. Baird et al., (2008) provide detail on the work of the Kenya Life Panel Survey, which has successfully tracked adolescents (a particularly hard age group to track) from 1998 to 2011 as they completed their education, married, and moved into the workforce. In the first round, 19% had

moved out of the district, and the team tracked respondents across Kenya as well as in Uganda, Tanzania, and even the UK.

3. **Think carefully about the timing of data collection**

   People are more or less willing or able to talk to enumerators depending on the time of day or year. Turn up in the middle of a work day and most people will not be at home. Call during dinner and they may not want to talk. Choosing the right time to collect data requires knowing your population well. It may also require paying enumerators extra to work outside normal working hours. Studies done at schools or workplaces have the advantage of keeping attrition down at relatively low cost as respondents are conveniently brought together in one location at specific times. Late afternoon or evening, when people have returned from work, is often a good time to interview people at their home. In rural Sierra Leone, enumerators stay in communities during surveys. This allows them to warn people the night before that they will want to interview them and arrange a mutually convenient time. It also means they are in the community at all times of the day, making it easier to find a time when people can be reached.

   Usually it is good to avoid doing a survey in traditionally high-travel months. August would be a terrible time to interview professionals in Paris, for example. The exception is if the study is tracking adolescents who may return to their parents' house during specific periods, such as Thanksgiving in the United States. When trying to track girls for our study in Bangladesh, we reduced our attrition by having a final round during Eid, a time when girls who are working in factories in Dhaka or have left for marriage traditionally return to their parents' houses.

4. **Collect tracking data at baseline**

   The baseline questionnaire should include a "tracking module" which asks questions like, "If you moved, who in the local community would know where you moved to, but would not move with you?" The tracking module should ask for phone numbers of the respondent and their relations.

5. **Can data be collected from people other than the participant?**

   Even if people have moved, or children have dropped out of school, it may be possible to collect some data on them from others who know them, which will minimize the costs of tracking and reduce attrition. Schools may know when a child dropped out. A child's peers may know if a girl got pregnant even if they are not still in school. Clinics may have data on when a patient stopped collecting their medicine and reporting for regular checkups (but note that the respondent's permission to get these data must be collected at baseline). Parents may know a lot about their grown children and parents often move less than their children.

6. **Make the survey as costless to answer as possible**

   Long surveys that ask stressful questions are likely to get lower response rates. The appropriate length depends on the respondent and means of data collection. Even if

respondents finish the baseline survey, they may deliberately make sure they are out in subsequent rounds if the survey is too long. Children have shorter attention spans so need shorter surveys. Phone surveys also need to be shorter than in-person surveys. If there are questions that might prompt someone to end the interview, such as questions on spousal abuse, these should be put at the end of the survey so that if the interview is terminated, only a limited amount of data are lost.

7. **Specify targets on attrition, not on the number of attempts made**

   It is common to specify the number of times an enumerator should attempt to reach a given respondent, but this can set up inefficient incentives. An enumerator has private information about when it is best to return to a household to maximize the chance of reaching the respondent, and it is important for them to have an incentive to utilize this (without having such a strong incentive to reduce attrition that they will fake data). Consider a phone survey where an enumerator has been given a list of people to call and told to call each at least three times. The easiest way to reach this goal is to call at a time when it is unlikely the person will be in and then call three times immediately one after each other. Attrition would be terrible in this scenario. If the enumerator is given a list of names and told to do what they can to reach as many as possible, they will learn about what times of the day seem to get high response rates, will ask when people might be available, and try the same person at different times of day.

8. **Consider compensation**

   Surveys take a lot of time and it may be appropriate to compensate people for this time. If the survey is long or a respondent needs to travel to a clinic or testing center to complete it, a small incentive may be useful in reducing the attrition rate. This is particularly true for panel surveys where the respondent knows that the survey will be long. Any incentive needs to be cleared with the institutional review board (IRB), which assesses the ethics of the study to ensure that people are not taking untoward risks because of the incentive. Compensating people for their time is usually seen as ethical by IRBs. Incentives that have been used include small backpacks for children, bars of soap, and seasoning cubes for cooking.

2.3 Poor data quality

The challenges of collecting high-quality data are not unique to field experiments, and some field experiments rely on administrative data (JPAL provides a useful guide to using administrative data in field experiments https://www.povertyactionlab.org/admindata). However, administrative data are often unavailable to researchers, not collected on all individuals, are unreliable, or not detailed enough for the researcher's needs, forcing the researcher to collect their own data. Data collection is hard and difficult to monitor which means enumerators can be tempted to take shortcuts and, in the extreme, make up data.

An essential part of data collection is therefore monitoring the quality of data, and critical to this is the back-check process. A highly skilled enumerator (usually a

supervisor) goes back and reasks a few questions from a randomly chosen subsample of respondents. The consistency between the two responses is then assessed. Many researchers do "back checks" of this kind on 10% of the sample. Because we need enough data to make this a valid comparison for larger surveys, the rate can be lower than this, and for smaller surveys it should be higher. The back-check survey does not have to be comprehensive. Indeed, the back check should be kept short to avoid respondents becoming annoyed at being asked the same questions twice. One reason for the back check is to make sure the enumerator is not making up data. Asking whether the respondent has been interviewed recently, and asking simple questions to which the respondent's answer is unlikely to change in the space of a few days are useful for achieving this. Enumerators should be warned that back checks will take place on an unannounced basis. It is good practice to make sure all enumerators have their work back checked at least once in the first few days of a survey and to discuss any important discrepancies between the two surveys with the enumerator. It should not be assumed, however, that all discrepancies are the fault of the enumerator. Respondents will often change their response depending on the day and how they are feeling, even when they are asked about slow-moving variables such as age or size of household.

Technology is providing an increasing range of options for monitoring enumerators. With paper-based surveys, monitoring has to rely on surprise visits from external monitors to check that the enumerator is in the right place at the right time. The monitor can also observe part of the interview to see if the enumerator is asking the questions well and appropriately recording answers. Paper checks can also be done: the team supervisor can pick up if certain questions are being missed or if a given enumerator has a high rate of failing to find target households. With GPS devices, enumerators can be tracked more closely. Even if we do not need to have the GPS coordinates of the interviewed household for the analysis, having enumerators record it helps ensure that they actually visited the household. Electronic data collection now allows part of or complete interviews to be recorded. Unlike having a supervisor listening over their shoulder, an enumerator does not know when the recording is on or which part of the interview recording will be checked, providing added motivation to perform well.

Electronic data collection also allows incoming data to be assessed while the survey is still in the field. By looking for patterns it is possible to find and correct errors enumerators may be making, and in worse-case scenarios to terminate employment. Warning signs include high variation between the answers collected by back checkers and enumerators, high rates of failing to find target households, and lower than average duration of interviews (measured by comparing the recorded start and end time of the interview). Surveys usually have important trigger questions in a survey which, depending on the respondent's answer, can change the survey's length. In a demographic survey there will be many questions for each pregnancy a woman has had; in an agricultural survey there will be lots of questions for each crop a famer grows. Enumerators who want to

keep their workload down have an incentive to have respondents answer a smaller number to these key trigger questions. Checking to see if certain enumerators have lower than average responses to these trigger questions is a good way to spot poor-quality enumerators. These trigger questions are also important to check during the back-check process.

Back checking is not able to solve the problem of respondents, not understanding the question, systematic under- or overreporting (which may be the result of, for example, social desirability bias), not knowing the answer, or being tired and inaccurate. Many of the chapters in this book discuss good practice in measurement. But it is also important to do extensive field testing in a given location with a survey instrument because questions that work with one population may not be well understood by another. It may also be necessary to develop locally relevant indicators especially for hard to measure and culturally specific outcomes such as social capital. Prior to the launch of the baseline survey for and evaluation of CDD program in Sierra Leone, Casey spent a year working with local partners to develop locally relevant indicators of collective action, trust, and participation (Casey et al., 2012). There is a tension, however, between relying on locally relevant indicators and internationally recognized indicators that can be used to benchmark levels and impacts across countries. If every study uses a different way of measuring outcomes it is hard to compare cost-effectiveness across projects because there is no single standard of effectiveness. Therefore, it is usually a good idea to have a mix of locally tailored and internationally recognized indicators. For example, in education studies we will want a test of learning that is appropriate to the level of learning in the population where the experiment takes place. However, if we are to compare program effectiveness across sites we want to also include some benchmark questions that can be compared across studies. For more discussion, see the chapter "Field Experiments in Education in the Developing Countries" by Muralidharan (2017).

2.4 Avoiding systematic differences in data collection between treatment and comparison

Most measurement issues that a researcher conducting an RCT has to deal with are similar to those faced by researchers working on studies using other methodologies. There are, however, a few issues that an RCT researcher has to be particularly concerned about. All of these boil down to the need for data to be collected in the same way in the treatment and comparison group and to avoid the intervention interacting with the way people report data.

Programs often collect a lot of data as part of their regular monitoring processes. These monitoring data can be very useful for interpreting the results of an RCT. For example, they can help us distinguish whether a null effect was due to a poorly implemented program or due to little impact from a well-implemented program. However, these program data should usually not be used to measure outcomes. If the program is operating only in the treatment area then there is no process data in the comparison areas, making a comparison

impossible. If we use program process data in the treatment area and try to collect similar data in the comparison areas, we will never know if any difference in measured outcomes is due to a real underlying difference in outcomes or due to a difference in measurement processes in treatment and comparison. For example, if data are collected by program staff in treatment areas and by professional enumerators in comparison areas, there is a risk that professional enumerators are better at probing respondents and checking inconsistent answers, and thus end up with systematically different outcomes than program staff.

In general, using program staff to collect outcome data is problematic as it can accentuate the risk of social desirability bias. Respondents may, for example, find it particularly awkward to admit to having practiced unsafe sex when asked by the person who trained them in the dangers of unsafe sexual practices. Data collection is also hard to do well, and there are considerable benefits from having it conducted by people who are highly experienced and motivated to do a good job because their future career prospects rely on them performing the tasks well.

The one exception where process data are sometimes used to measure outcomes is when the RCT takes place within a sample in which everyone participates in the program, the randomization is into different types of program participation, and process data are collected routinely on those in treatment and comparison in identical ways. For example, if different borrowers within the same credit organization are randomized to receive alternative versions of the credit contract and repayment is the outcome of interest, then the lender's information on repayment rates can be used to compare outcomes for treatment and comparison clients (Giné and Karlan, 2014) use this approach when looking at microcredit contracts, and William et al. (2016),[9] use this to look at farming cooperative contracts—although both collect survey data as well. Even in these cases, it is useful to check the validity of the data by comparing self-reported data from surveys with administrative data from the implementing organization, especially if there is subjectivity in the measurement of outcomes. The concern is that to the extent that program staff are collecting process data and know which participants have been allocated to treatment and which to comparison, this knowledge and any biases they have about outcomes may influence how they record outcomes.

Another temptation is to collect data on the treatment group at a different time than the comparison group. For example, if the partner is pushing to get the program implemented quickly they may request that baseline data are collected in the treatment area first so that the program can start, with data collection done in comparison areas later. This timing difference compromises the difference between treatment and comparison data and should be avoided.

[9] A description of this ongoing study can be found at: http://www.povertyactionlab.org/evaluation/encouraging-adoption-rainwater-harvesting-tanks-through-collateralized-loans-kenya.

If the program has an impact on the relationship between the underlying outcome and the measurement of the outcome—even if data are collected in the same way in treatment and comparison—the data cannot be interpreted the same way in the two groups, thus undermining the validity of the experiment.

This problem most often arises when a program provides an incentive to change a particular behavior, which also changes the incentive to misreport the behavior. We want to be able to distinguish between the incentive leading to changes in actual behavior and the incentive leading to changes in reported behavior but not actual behavior. The more objective the measurement of the outcome, the less likely this is to happen, but if the incentive is high enough it is possible that it will induce substantial cheating that can corrupt even more objective measures. This is why it is preferable to use an outcome measure separate from the measure that is used for the incentive. For example, Dhaliwal and Hanna (2014) study a program in which medical worker attendance is monitored with a threat from officials that action will be taken against those with high absence rates. To judge if the program impacted attendance, the authors use random checks that are not linked to the official monitoring. Even if a program does not change respondents' incentives to report an outcome, it may change the perceived social desirability of a behavior. For example, a program designed to encourage saving may make people more liable to report saving even if it does not change saving itself. In situations where the program may change social desirability, it is imperative to rely on more objective measures of outcome, often including nonsurvey outcomes. Glennerster and Takavarasha (2013) have a catalog of nonsurvey outcomes with the pros and cons of each.

3. ETHICS[10]

Most field experiments involve humans as subjects in their research, and in this they are no different from most empirical economic research. But the expansion in the use of field experiments has been associated with more researchers, and more junior researchers, collecting their own data, especially in developing countries. There are a host of practical challenges associated with collecting and storing confidential data, which we discuss in this section. While most of the practical and ethical issues involved in running field experiments are common across any research that involves primary data collection, the intense collaboration between researchers and implementers common in field experiments does raise specific ethical questions, particularly in relation to the boundary between practice (which is regulated by national laws as well as norms and professional ethical standards) and research (which in most countries has separate formal regulatory structures).

[10] This section draws on Glennerster and Powers in The Oxford Handbook of Professional Economic Ethics, edited by George DeMartino and Deirdre N. McCloskey (2016).

The basic principles underlying the US system of ethical research regulation were set out in the Belmont Report. This report was issued in 1978 by the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research and provides the basis for decisions about the ethics of research funded by most federal departments or agencies (Code of Federal Regulations, title 45, sec. 46.101).[11] While the principles set out in the report were formulated in the United States, they are reasonably general and are similar to the principles behind institutional review structures around the world.[12] Since 1978 hundreds of thousands of research studies have been evaluated against these principles building up a considerable bank of experience in how to apply them in practice.[13] The principles explicitly cover both medical and nonmedical studies and recognize that the level of review and safeguards should be adapted to the level of risk for a given study. This is important as social science research often has lower levels of risk than many medical studies.

There are three key principles spelled out in the Belmont Report:

1. **Respect for persons**

   People should be treated as autonomous agents. They have their own goals and have the right and ability to decide the best way to pursue them. In most cases this principle requires that researchers clearly lay out the risks and benefits of the study to potential participants and let them decide if they want to participate. The principle also recognizes that there are individuals who do not have full autonomy, such as children who may not understand the full risks and benefits of the research, or prisoners, who may not have freedom of action. Where autonomy is compromised, the researcher has to take special precautions.

2. **Beneficence**

   Researchers should avoid knowingly doing harm and seek to maximize the benefits and minimize the risks to subjects from research. However, avoiding all risk of harm is unrealistic and would prevent the gains to society that come from research. Therefore, risk of harm needs to be weighed against likely benefits to society that could flow from the research.

3. **Justice**

   The justice principle focuses on the distribution of costs and benefits of research. It seeks to avoid a situation where one group of people (for example, the poor or prisoners) bears the risks associated with research while another group receives the benefits. It recognizes that the individuals who take on the risks of research may not be

[11] Accessed at http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.101, August 15, 2013.

[12] For example, the Australian guidelines similarly include principles of justice, beneficence, and respect, although they also include a "research merit and integrity" principle. The three main principles underlying Canadian ethics review are respect for persons, concern for welfare, and justice.

[13] PubMed, a database of medical research, reports over 325,000 medical trials registered between 1978 and 2013.

precisely those who reap the benefits. Instead it aims to ensure that research is conducted among the types of people who will benefit from it.

The principles are a compromise between two somewhat separate ethical traditions: a rights-based approach and a utilitarian approach. The beneficence principle's emphasis on the need to weigh risks (which fall on the individual) and benefits (many of which accrue to society) is familiar to utilitarians and economists. It is modified by the right to self-determination in the respect-for-persons principle: Research that imposes risks on the individual for the sake of society is ethical, but only if the individual understands the risks and is willing to take them. But the right to be informed from the respect-for-persons principle is not absolute and is itself modified by the beneficence principle: Where the risks associated with the research are minimal and the costs of fully informing the subject are large, it is ethical to not fully inform, and in some cases even deceive, subjects. The costs in this case can be monetary or costs to the effectiveness of the research.

The justice principle explicitly addresses one of the objections to utilitarianism—that it justifies harm to some if it creates benefits to others—by saying that those who take the risks should receive the benefits. But by applying the principle to groups of people rather than individuals, it is a compromise between the two ethical traditions.

3.1 Institutional review boards

As the principles make clear, there are difficult trade-offs to make when determining the most ethical way to proceed with research. Researchers have the primary responsibility for judging these trade-offs. However, they also have an interest in moving ahead with their research, which may blur their perceptions of risks and benefits. An independent authority is therefore needed to assess the trade-offs and ensure that ethical rules are applied appropriately. IRBs fulfill this role. Most universities in the United States have IRBs with their own processes for reviewing and approving research conducted by faculty, staff, and students at the university. Research funding from most agencies of the US government requires that researchers follow a set of ethical review guidelines established by the Office for Human Research Protections (OHRP), and these guidelines have therefore become the default standard applied by universities even when a study is not funded by the US government. OHRP standards flow from the Belmont Report but are updated regularly.[14]

Some US nonuniversity research organizations maintain their own internal IRBs, which follow OHRP standards (for example, Innovations for Poverty Action and Abt Associates). Others, such as Mathematica Policy Research, use external IRBs accredited

[14] Available at http://www.hhs.gov/ohrp/humansubjects/commonrule/index.html. See also http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.101.

by the Association for the Accreditation of Human Research Protection Programs, a voluntary organization.

Outside the United States, the system of ethical review for social science research that involves human subjects is quite mixed. Some countries have systems similar to the United States. Australian research guidelines, for example, include principles of justice, beneficence, and respect, although they also include a "research merit and integrity" principle (which in the Unites States is integrated into the beneficence principle). The three main principles underlying Canadian ethics review are respect for persons, concern for welfare, and justice.

A surprising number of universities outside the United States have no formal system of ethical review for research involving human subjects. Because ethical review boards have mainly been seen as the province of medical research, many universities that do not have medical schools do not have ethical research review boards. In addition, some medical review boards either do not accept nonmedical research for review or are ill equipped to review nonmedical studies.

Social scientists face three main problems when seeking review from medical review boards: these boards are unfamiliar with the type of work social scientists undertake; they have procedures that are designed for studies that impose much higher risks on subjects; and they impose medical ethics standards, which are not the same as research ethics standards. Lack of familiarity can mean that questions are raised about outcome measures that are standard in social science (I was once asked to remove a question about what assets a household owned from a survey as it was seen as too intrusive). Because medical boards are used to dealing with studies that impose substantial risks on subjects, they often have more rigorous safeguards as standard requirements than is normal in low-risk social science studies and are unwilling to approve waivers for informed consent or written informed consent, even when the risks are low and the burden very high. If a study is examining the impact of a new drug that may have dangerous side effects, it is probably appropriate to get written consent from illiterate participants by having someone they know carefully read to them the consent form that lists all the risks and have them sign. If the study simply measures their height, it may still be regarded as a "health" study but gaining oral consent from illiterate participants should be justifiable. Doctors and nurses have ethical obligations that go beyond research ethics including providing care to those in need. Thus medical ethics boards may require researchers to offer medical care to those they find are in need of care as a result of their research. For example, if anthropometric measurements reveal that a child is malnourished, a medic may be expected to refer them to care. While medical boards may require treatment of subjects that researchers find to be ill, this obligation does not flow from most research ethics principles.

Some researchers working on field experiments have responded to the lack of IRBs by working with their universities to establish such review boards. The Paris School of

Economics and the Institute for Financial Management and Research in India worked with J-PAL Europe and J-PAL South Asia, respectively, to establish IRBs in 2009. The World Bank, which currently relies on the regulations in its member countries, is actively discussing the creation of an ethical review board (Alderman et al., 2014). It is somewhat surprising that the field experiment movement should have spurred the creation of IRBs as many of these institutions (including the World Bank) collected data from human subjects long before field experiments became popular.[15]

3.2 When is ethical review required?

Researchers have to seek ethical review when they conduct research that involves human subjects. The precise definitions of "research" and "involving human subjects" can vary between jurisdictions, so a researcher needs to understand the local rules that apply to their research. In some cases multiple standards apply (for example, when a researcher at a US university conducts a study in Kenya, they may need to seek approval both from their home university and from the Kenyan Medical Research Institute (KEMRI).

In the United States, "research" is defined as systematic investigation that leads to the creation of general knowledge. Process data about the functioning of a program is not research because it is designed to inform the program, but not to generate general knowledge that is useful for other programs. Asking a few beneficiaries of a program about their experience is not research because it is not systematic (and therefore does not generate general knowledge). This is why most internal evaluations done by nongovernmental organizations and governments do not count as research and are not subject to the same rigorous ethical review.[16]

The practical implication of this definition for researchers is that the early stage work that researchers do to prepare for a field experiment does not usually count as research and thus can be done prior to ethical research approval. For example, researchers may visit the program and talk to beneficiaries and program staff. They may examine administrative data and pilot questionnaires, all before approval has been given. Indeed, much of this work is needed to prepare the paperwork for ethical review, as most reviews require a copy of the final questionnaire to be used in any primary data collection. Approval (or a waiver stating that full approval is not required) needs to be secured before the collection of any data that will be used in the study and that is collected for the purpose of the study. Data that are used in the study but are not collected for the purpose of

[15] One potential spur to the creation of IRBs is the relatively new requirement instituted by the American Economic Association that papers involving the collection of data on human subjects must disclose whether they have obtained IRB approval.

[16] This is the case even though internal evaluations often collect similar kinds of data to those collected in field experiments and the risks associated with inappropriate release of the data is similar. In some countries NGO or governmental handling of data from internal evaluations is covered by privacy regulations.

the study (including ongoing administrative data collection) can take place before approval has been received because it would have gone ahead with or without the study. However, approval may be required for the researcher to access and utilize even administrative records because these can include personal information, the release of which could cause harm to a research subject.

The second trigger for ethical review is that the research involves human subjects. (There are other guidelines for research on animals, but as social science rarely has animal subjects we ignore these regulations here.) Research counts as having human subjects if it includes interviews with human subjects, or collects physical specimens from humans (e.g., urine or blood).

If research involves use of data about humans but does not involve the collection of that data, and the researcher never has access to information that would allow them to personally identify them, then ethical approval is not required. Nor is approval required if the researcher only uses publicly available data (which usually has all personal identifying information removed before being made public). Thus a study which uses data from a Demographic and Health Survey would not require ethical approval. Much like the use of administrative data, if the researcher needs to acquire personal identifiers (such as precise geographic location) to undertake their research, then approval is required even if they do not collect the data themselves.

3.3 Practical issues in complying with respect-for-human-subjects requirements

1. Informed consent

The respect-for-persons principle requires that researchers explain any risks of harm associated with participating in the study to those involved and gain their consent before proceeding.

In the case of an experiment randomized at the individual level, complying with this requirement is usually relatively straightforward. We select the study sample and then approach the individual, inform them of any risks associated with participating in the study, and request their consent to participate. Usually this is done before randomization, in the context of collecting baseline data. If the subject does not consent they are dropped from the sample, although it is good practice to record the number of subjects who decline to participate to give a sense of the representativeness of those who do participate.[17] The precise wording of the consent and the method by which it is collected has to be approved by the IRB and depends on the circumstances of the experiment and the risk involved. In general, written consent

[17] Information on the number of those approached who declined to participate is a requirement under consortium guidelines, and thus usually has to be included in a paper published in a medical journal.

(i.e., having a subject sign a consent form which sets out the risks and any potential benefits) is preferred. However, when many of the subjects are illiterate, a written consent form may not be the most effective way to convey risks. It may even cause distress to ask illiterate subjects to place their mark on a written document they cannot read. Alderman et al. (2013) suggest that in India, asking an illiterate person to provide their mark on a paper as part of the interview process may give the impression that the survey is run by the government (as thumb prints are often associated with official documents) and that therefore participation is mandatory, undermining respect for persons. If the risks are high, we may nevertheless need to get written consent by finding a literate member of the community and trusted by the participant to carefully explain the written document to the participant. For the most part, however, social science experiments do not involve this high level of risk and gaining oral consent is often appropriate, especially when a high proportion of subjects are illiterate. In this case, the enumerator reads the consent language and asks if the subject provides consent, and then checks a box if this consent is given. A key part of consent language is explaining that the subject has the right to leave the experiment at any time and has the right not to answer any question during the data collection process. It is important that the consent is written in a way that subjects readily understand. Zywicki (2007) provides examples in which IRBs have made consent forms more technical and harder to understand—which makes it harder for those with limited education to make informed decisions about participation.

Collecting informed consent when randomization is at the community level is more complicated, as data are often only collected on a random sample of those in the community and thus the research team may not interact directly with all individuals in the community. There are three important issues to keep in mind when determining how to proceed in this situation: Does the program require participants to opt in? Will data be collected on community-level outcomes, in which case all members of the community are under some definitions subjects of the experiment? To what extent is the program itself standard practice, and thus those who participate in the program but from whom no data are collected are not considered part of the research?

Many of the programs that are evaluated by field experiments require participants to opt in. For example, if a program offers the chance for mothers in a given community to attend literacy classes, mothers have a chance to opt in or out of the treatment. As we discuss later, some IRBs would not consider those who take part in the program but on whom the researcher does not collect individually identifiable data, as being subjects of the experiment. However, even if these program participants are considered subjects of the experiment, the program is compliant with the principle of respect for persons if someone explains the program to potential participants, who then choose whether or not to participate.

The ethical issues become more complicated if the program provides a service to the entire community that participants cannot opt out of (Hutton, 2001). Examples include adding chlorine to the community well, erecting streetlights, modifying the rules under which the mayor is elected, or changing how teachers teach. Usually implementing organizations have ways of seeking community assent before proceeding with this type of community-level intervention, and are either governmental bodies themselves with their own processes of accountability, or are regulated by government as implementing bodies. If the risks of the intervention are low, then individual consent from all community members is not usually required: either because the IRB decides the costs of collecting it are too high given the small risks or because they consider the program implementation as practice rather than research and thus outside their purview. The exception might be if the program design were considered to be driven more by research considerations than program considerations (we discuss this issue in the next section).

In many medical clustered RCTs, informed consent is not collected from individuals because individuals are not considered the subjects of the trial, especially if the intervention works at the level of the medical practitioner. McRae et al. (2011) argues that patients are not the subjects of trials that provide different types of training or incentives to doctors. This is because researchers do not directly interact with patients, while medical professionals, who should be considered the subjects of the trial, are ethically responsible for deciding what is right for their patients.

2. Waiving informed consent

Research ethics rules allow the requirement for informed consent to be waived when the risks to the subject are low and the costs of collecting informed consent are high. The costs of collecting informed consent could be monetary or come in the form of damaging the integrity of the research. Imagine an experiment on the effectiveness of different forms of advertisement in reducing smoking amongst adults. The experiment randomizes the position of antismoking billboards across the United States and then measures the level of smoking from sales of cigarettes. The participants of the study include anyone who sees the billboard. The researcher has no good way of identifying the individuals who see the billboard, and data to assess the effectiveness of the intervention comes from administrative records on cigarette sales, so they have no opportunity to ask for consent during data collection. Going door to door in the area to collect consent would be prohibitively expensive and the risks of harm from seeing a billboard are low, so the research is likely to receive a waiver for informed consent. Similarly, many education field experiments in the United States are exempt from collecting consent from all parents of students, as it would be infeasible and the risks are low.

The other cost of collecting informed consent is that knowing they are part of a study, or knowing the full details of the study, could change a subject's behavior,

which could undermine the validity of an experiment. We may not want to tell people, for example, that they are involved in an experiment on racial bias as this may make them more aware of potential bias and thus change their behavior during the experiment. One approach is to tell the subject they are part of a study, but not give a full explanation about what the experiment is about, or even mislead the subject about what the experiment is about. Another approach is not to tell subjects they are part of an experiment. If we do not tell people they are part of an experiment or mislead them about what the experiment is about, permission is required from an IRB before the experiment can go forward. A researcher must justify the waiver of informed consent by explaining the likely benefit of the research to society, and why the research would be undermined if the subjects knew they were part of an experiment or knew the real reason for the experiment. The IRB will then decide if the lack of full transparency is warranted. IRBs will often require researchers to debrief subjects at the end of the experiment as a condition for gaining the waiver.

Note that this is different from deception within the experiment, which is when a research tells a subject something that is untrue as part of the experiment. Perhaps the most common form of deception in field experiments is when enumerators pretend to be someone they are not: for example, pretend to have a specific set of symptoms to see whether the medical professional asks them the appropriate questions and responds to the answers with the appropriate care recommendations. One way to achieve informed consent in these situations is to warn the provider in advance that there will be mystery patients at some point and get their consent for this test. If the experiment runs over many months, this knowledge that one of many patients will be a mystery patient is unlikely to dramatically change their behavior. For more on deception and informed consent, see Alderman et al. (2013).

3. Protecting confidentiality of information

As part of informed consent, the subject is usually told that any information they provide will be kept confidential. This agreement with the subject must be strictly adhered to and an IRB application needs to set out the practical steps a researcher will take to comply with this agreement. Anyone in the research team who is involved in handling data—from the enumerator to the principal investigator—must be trained on proper data handling to ensure that the protocols described to the IRB are followed. Important ways to ensure the maintenance of subject confidentiality are to ensure that any information that can link the data back to an individual (i.e., personal identifiers), such as name, address, phone number, or photo, is separated from the rest of the data as rapidly as possible; that only deidentified data be used, wherever possible, during analysis (to prevent the risk of data leaks); and that data with personal identifiers are kept secure. The precise steps will depend on what the data consist of and how they were collected. For example, when data are collected through paper surveys, all personal identifiers should be put on the first one or two pages of the

survey and an ID number (generated only for the purposes of the research and thus uninformative to anyone else) should be printed on all pages of the survey. This means that as soon as the survey is completed and checked by a supervisor in the field, the first pages with identifying information can be separated and stored separately from the rest of the survey. The pages with the identifying information and the codes that link that back to the answers to the survey must then be stored in a secure place (such as a locked cabinet). When data are collected electronically, the device can be encrypted so that if the phone, tablet, or PDA is stolen no one can access the data. If analysis does require some identifying information (for example, global positioning data to examine geographic spillovers), the analysis needs to take place on an encrypted computer so that if the computer is stolen the data cannot be accessed. As we discuss later, when identifying information, such as global positioning data, is an integral part of the analysis, it can be complicated to publish sufficient data to fully replicate the study while still maintaining confidentiality.

3.4 The ethics of implementation

In the discussion of informed consent, it became apparent that it is not always straightforward to identify who is the subject of research and thus from whom informed consent is required. In particular, when a field experiment is evaluating a program, are those involved in the program but on whom the researcher does not collect data, subjects of the research or not? and do research ethics thus govern the program? The Belmont Report notes that the line between research and practice, and thus the line between what requires ethical approval and what does not, is blurred. While most of the report is appropriate both for biomedical and behavioral (or social science) research, the section that deals with the distinction between research and practice is written almost entirely from a biomedical perspective. This has led to some confusion and debate about the ethical standards to be applied to the implementation of programs that goes alongside many social science field experiments. Indeed, the Belmont Report explicitly states, at the end of the section defining the separation of research and practice, that the authors do not feel equipped to define the boundary between research and practice in social science:

> *Because the problems related to social experimentation may differ substantially from those of biomedical and behavioral research, the Commission specifically declines to make any policy determination regarding such research at this time. Rather, the Commission believes that the problem ought to be addressed by one of its successor bodies.*

Subsequently, a group was established to work on this, but no additional guidelines were released. The practical question that faces researchers and IRBs evaluating research proposals from social scientists is if and when ethical approval should be sought, and research rules (including requirements for informed consent) applied to the program

that is being evaluated. The discussion in following paragraphs represents my view based on a close reading of the Belmont Report and requesting ethical review for many RCTs. However, it is worth reiterating that different IRBs in the United States interpret the standards differently; different countries have different rules; and the regulation of implementation is one of the areas where standards differ most sharply across institutions.

At one end of the spectrum the answer seems obvious: in the canonical case of a medical field experiment testing a new drug, the risks associated with the drug (the intervention) need to be assessed against the benefits of learning about its effectiveness. In other words, the assessment of risks and benefits and the informed consent apply to the program being tested (the drug) as well as the data collection that surrounds it.

Yet there are also examples where it is equally obvious that ethical regulations have no jurisdiction over the intervention a researcher is evaluating. Angrist (1990) evaluates the impact of the Vietnam War, which involved a lottery to determine participation. Chattopadhyay and Duflo (2004) similarly evaluated the impact of a ruling by the Indian Supreme Court that the position of village leader (*pradhan*) had to be given to a woman in one-third of cases (allocated randomly in many Indian states). In these cases IRBs had no jurisdiction over the implementation of the program being evaluated: there was no question of insisting that those whose names were entered into the Vietnam lottery had to provide informed consent. Nor could villages decline to participate in the quota program for women's political participation.

What is the key distinction between the evaluation of a new drug and the evaluation of the Vietnam War/quotas cases that explains why implementation is part of research for the first, but not the other cases? One difference is that the drug (the intervention) was designed by the researcher, whereas in the other two cases the intervention was designed and implemented by someone else (e.g., the government or the Supreme Court). I do not think this is the *key* distinction for two reasons. First, we think the review of the drug trial should include the risks and benefits of the drug whether or not the researcher who developed the drug goes on to test it, or if someone else runs the clinical trial. Second, if the identity of the implementer determines whether the intervention should be reviewed, then we would say that if a researcher also helped run a nongovernmental organization, then everything that NGO did, whether or not it was evaluated, should be subject to ethical approval.

The Belmont Report also supports the idea that whether or not an activity falls under research guidelines should be based on what the activity is, not on who undertakes it. The report acknowledges that (for biomedical research) researchers will often practice medicine (just as social science researchers sometimes practice direct poverty-alleviation work or advise governments or NGOs on the design of policy). This "practice" is deemed to fall outside the purview of research ethics. Instead, the Belmont Report defines research as an activity that leads to generalizable knowledge.

The challenge in applying this rule in the case of field experiments is that it is a combination of two different activities that lead to generalizable knowledge. Most field experiments combine the rollout of a program with data collection, and neither on their own would create generalizable knowledge.

But this gives a useful criterion for deriving whether and what part of implementation falls under research ethics guidelines: namely any change in program implementation from normal practice (or what would have happened otherwise) that is brought about for the purpose of creating generalized knowledge. Thus if a program was to be rolled out by an NGO in a new area anyway, this would not create generalized knowledge and would not (in my view) count as research, and the program itself should be governed by the regulation of NGO activity rather than a research ethics board. The use and collection of data by researchers studying the rollout does fall under research ethics, as it is necessary to draw general lessons from the rollout. However, if to learn from the program the rollout was changed in a substantive way, then this change is covered by research ethics. Note that this is not the position that all IRBs take. KEMRI required that parents of all children who were part of a school-based deworming program in Kenya run by International Child Support provide written permission before receiving the drug because the program was being studied. If the program was not being evaluated, the NGO would not have had to collect written (or even oral) consent as deworming drugs have been shown to be extremely safe. In other words, exactly the same action by the same organization was considered research when the action was being studied, but was not considered research when not studied. Zywicki (2007) discusses an example where a study that included provision of a potentially life-saving medication was shut down because researchers were unable to get signed consent in advance—even though in the absence of a study, written consent would probably not have been required to provide the medication.

It is sometimes assumed that if a researcher implements a program, then the entire program is part of the change that is introduced with the purpose of generating knowledge. But as I have argued earlier, ethics guidelines are not based on who does the activity, but what the activity is. Thus if a researcher evaluated an NGO program that hands out bed nets at a school and the researcher interviews a random subset of children at the school, then the researcher would only have to get informed consent from the individuals who they interview. If the researcher organization is the one to hand out the bed nets, I would argue the same rules apply: research rules cover the interviews and data collection, but informed consent is not required to hand out the bed nets themselves.

Questions around researcher implementation were vividly illustrated in a controversy surrounding an election experiment conducted in Montana (Johnson, 2015). In the experiment, researchers sent voters flyers that put individual judges up for election on an ideological scale. Key complaints about the project were that (1) the flyers used the State of Montana seal, giving the impression that the document was an official state

document when it was not; (2) the flyers were "express advocacy," i.e., they advocated for individual candidates rather than issues and thus fell underreporting rules which were not followed; (3) IRB approval for the study as it was carried out was not sought or received; and (4) the intervention may have changed the results of some elections. The first two are violations under Montana election law according to the report of the Commissioner of Political Practices of the State of Montana (2015) and are being dealt with as such (Motl, 2014). In other words, the researchers are being regulated as implementers and being held to those standards. This fact adds one twist: due to universities' tax status, even if the researchers had followed disclosure rules to express advocacy, any money that ran through the university could not be used for advocacy.

There is more debate about whether changing the outcome of an election is a violation of ethics. Presumably the objection applies only if researchers run the intervention, because researchers study interventions that influence elections all the time. If the view is that interventions run by researchers should not change elections, this raises the question of whether interventions run by researchers should not change other outcomes. It would be odd to say that we do not want field experiments in medicine to change peoples' health outcomes, for example.

One argument to suggest that elections are different from other interventions is that while improving one person's health does not influence another person's health, election outcomes are a zero sum gain; an intervention cannot contribute to an overall improvement in society and instead must inevitably help one group at the expense of another. But many of the interventions that researchers study have some distributional or zero sum aspects. Is it unethical for a researcher to run a study that helps some women establish small businesses, which could have a potential negative externality on existing local businesses? The truth is that social science is involved in the real world and the interventions that social scientists study will have impacts in that world. One practical call for change that has come out of the Montana case is that IRBs may be too focused on potential harm to the narrow subjects of the experiment and should be more aware of costs to society as a whole, as well as benefits to society as a whole (Humphreys, 2014; Johnson, 2015). As we conduct studies we must be aware both of research ethics and the ethics and regulations surrounding the interventions we study. But it is unclear why researchers should, when acting as implementers, have a different set of ethics standards or regulations from other implementers.

One benefit of deciding what should be covered by research ethics based on the activity and not on who undertakes the activity is that it avoids drawing a bright line about when a program is researcher implemented and when it is not. Given the close partnership between researchers and implementers in field experiments, most programs that are evaluated are a combination of the two. Even when someone who is not a researcher implements a program, the researcher often provides advice (based on their knowledge about what has worked elsewhere) about the program design. But advice about how to

improve a program is not research. What counts as research is deliberately manipulating the program to produce general lessons: for example, to create a control group so that the program can be evaluated rigorously. In the next section we discuss examples of where there might be potential risks or costs associated with the changes in implementation brought about by the manipulation of a program necessary to rigorously evaluate it.

3.5 Potential harm from different forms of randomization

There are many different ways of introducing an element of randomization into a program to enable rigorous evaluation of its impact. Each approach raises its own unique ethical issues.

The research manipulation that nonresearchers often feel most uncomfortable with is the treatment lottery. In this design, some study participants are randomized to have access to the program and some never receive the program. The concern is that some potential participants in a program are "denied" access to the program in order to evaluate its impact. When assessing potential harm from a field experiment we need to consider whether the introduction of a treatment lottery changed the total number of people who receive the program or whether it changed who received the program. In most cases, the treatment lottery approach is used when there are insufficient funds to provide the policy or program to all those who could benefit from it. For example, a program provides financial literacy training to small-scale entrepreneurs in Bolivia but only has funding to cover 200 entrepreneurs, far fewer than the number of all eligible entrepreneurs. A lottery is used to decide who receives access to the program but does not change the number of people treated.

There may be cases where a program (often a government program) does have sufficient funds to provide the treatment to all those who are eligible, but a decision is made to reduce the number of people who receive the program in the first phase to evaluate it. In this case, the risk of harm is that the program is beneficial and delaying its introduction to all of the eligible delays benefits to those potential participants. Note that this is a risk of harm, not a known harm, because at this stage we do not know that the program will be beneficial. (If we did know it was beneficial and there was funding for everyone to receive it, we should not be doing the experiment.) This risk of harm needs to be offset against the potential benefits of understanding the impact of the program, including the possibility that we find the program has unanticipated negative effects and that evaluating it saves people from these harms.

If a treatment lottery does not change the number of participants in a program, it might change who participates in a program. Ravallion (2012) suggests that allocating benefits randomly treats research subjects "merely as means to some end," and thus violates the respect-for-persons principle. But all research with human subjects uses information from some individuals as a mean to the end of drawing general lessons. Especially if

the risk of harm is small (for example, the time cost of filling in a survey), and even when they are large (as in some medical trials), many people are happy to contribute if they feel there are benefits from the research to society.[18] The respect-for-persons principle recognizes that people can make informed choices about whether to participate in a study that may mainly help others.

A subtler objection is that random allocation of resources is a form of mistargeting (Barrett and Carter, 2014). Imagine that a program has funds to provide warm clothing to 500 poor families in a city in the northeastern US, and the implementers have a good way to identify those most in need. Evaluating this program would require identifying 1000 needy families, some of who might not be as needy as the original 500 if the program had really identified the 500 neediest families in the city. From the 1000, half would be randomly chosen to receive the warm clothing. In this case, the evaluation imposes some risk of harm because some of those identified as the 500 most needy will end up not receiving the warm clothes, while some who are slightly less needy will receive them. Note, however, that it is only a *risk* of harm because we do not know if receiving the warm clothes is a benefit (if we did we would not be evaluating the program) and we usually do not know whether the way that the program identifies the neediest is effective. Recent field experiments that specifically look at the question of targeting (by randomizing different approaches to targeting in different communities) suggest that conventionally used targeting approaches may not necessarily be the best way to identify need (Alatas et al., 2013). Many programs do not do a comprehensive assessment of who are the neediest in a given target area. Instead they have eligibility criteria and stop recruiting to their program when it is full. In these cases, it is possible to work with implementers to continue the recruitment process until a larger number of eligible participants have been identified and then randomized among them. As the most vulnerable are often not the first to sign up to a new program, this extended recruitment period can actually help improve targeting.

When designing a field experiment it is usually possible to avoid weakening the targeting criteria of the program by expanding the geographic scope of the program. In the example stated previously, instead of expanding the potential pool of families to 1000 in the same city, it might be possible to expand the program to a second city, identify the 500 neediest families in each, and then randomly pick 250 from each to receive the program. This would allow the evaluation to go ahead without weakening the targeting. This geographic expansion to accommodate an evaluation does usually increase the logistical costs of the program implementers, and this cost needs to be set against the benefit of doing the evaluation.

[18] As we discuss under the respect-for-person principle, there is often a challenge of getting informed consent in clustered trials.

If none of these options are workable and there is a high risk that the evaluation will lead to poorer targeting of the program, this would not necessarily make the evaluation unethical, because this risk needs to be compared to the benefits associated with the study.

One form of field experiment where the issue of mistargeting is particularly relevant is the treatment lottery around a cutoff. Unlike a simple treatment lottery, this methodology explicitly recognizes that some potential participants may be more qualified than others and is used when programs have explicit criteria for ranking eligibility. Potential participants who are near the cutoff for eligibility are randomized into or out of the program. There are three slightly different ways to do a lottery around a cutoff. Eligibility can be expanded to those who would previously have been ineligible, and access to the program within this group can be randomized. Or the group that is to be randomized can come out of those who would previously have been just above and just below the eligibility cutoff. Or the randomization can occur only among those who would previously have been eligible, thus reducing the total access to the program. Usually the methodology does not change the number of beneficiaries, but in most cases it involves accepting some people into the program who are less qualified than some others who are not accepted.

In assessing the trade-off between costs and benefits of using a lottery around the cutoff, there are a number of issues to keep in mind. As we have said, it is unlikely that the program is known to be beneficial, or else the evaluation would not be occurring. There are degrees of uncertainty: the stronger the evidence that the program is beneficial, the greater the concern about "denying" people access. Another key question is whether the benefits of the program are likely to be higher for those who appear to be more qualified.

For example, imagine the methodology is being used to evaluate the effect of giving access to consumer loans to people in South Africa (Karlan and Zinman, 2010). The bank has a scoring system for deciding who is creditworthy. The assumption is that those who score highly will use the loan wisely and will be able to repay the bank, making both the bank and the participants better off. The scoring system is also meant to weed out those who would be a bad risk and will not be able to repay. Potentially bad risks do worse if they are given a loan and cannot repay it because they acquire a bad credit record (although if they would never otherwise have been eligible for a loan from any lender it is not clear a poor credit record hurts them). It was precisely this concern on the part of the bank about the quality of their targeting approach that led them to invite the researchers to study the cutoff and help them improve it.

But do the researchers, or the bank, know that the scoring system is good at determining who is a good risk and who is a bad risk? Maybe the system is good enough to detect the very good and the very bad risks, but does it do a good job of selecting people around the cutoff? It is also possible that the credit scoring system may be discriminating against people who are good risks but happen to live in a poorer

neighborhood. In this case, using a lottery may actually reduce the harm of discrimination. If there is uncertainty about the quality of the scoring system, a lottery around the cutoff can be a very good reason to do a randomized evaluation, because it helps generate knowledge about how good the scoring system is and whether the cutoff has been placed at the right point. It was precisely this uncertainty about the appropriate scoring system and cutoff that led the South African bank in Karlan and Zinman (2010) to ask the researchers to undertake the research.

In the bank example, if the evaluation finds that those just below the cutoff do just as well as those above it, then the bank will be encouraged to extend its loans to more people, and those just below the cutoff will gain, as will the bank. There is a risk that the cutoff was at the right place and that those below the cutoff will get into debt as a result of being offered a loan they cannot repay. This risk has to be taken into account when designing the study. The risk can be ameliorated by only randomizing above the cutoff (lottery among the qualified) but this has other risks: the evaluation cannot tell if the cutoff was too high, and it reduces access among the qualified more than in other designs. It is also possible to narrow the range around the cutoff within which the randomization takes place so that the bank never lends to anyone who has a very bad score. But this also has downsides: less would be learned about where the cutoff should be and, for a given size program, there would be less statistical power and thus less precision in the impact estimate.

The better the evidence there is that the cutoff is well measured and targets the program well, the more careful researchers should be with a lottery around the cutoff. For example, there is a strong evidence base suggesting that weight-for-age and arm circumference are good criteria for judging which children need a supplemental feeding program. Researchers may therefore decide that randomizing around the cutoff for a supplemental feeding program is not appropriate.

4. TRANSPARENCY OF RESEARCH[19]

Organized skepticism is essential to the process of scientific inquiry: "Involving as it does the verifiability of results, scientific research is under the exacting scrutiny of fellow experts. … The activities of scientists are subject to rigorous policing, to a degree perhaps unparalleled in any other field of activity" (Merton, 1942, p. 276, as quoted in Miguel, 2015).

In the last few years there has been growing concern that research in the social and medical sciences does not always live up to this ideal. In 2011 an investigation of the

[19] In preparing this section, I learned a lot from the lecture notes of Edward Miguel's semester long course on transparency of research (http://emiguel.econ.berkeley.edu/teaching/12), although I do not always come to the same conclusions.

work of Diederik Stapel revealed at least 30 papers in peer-review psychology journals were based on made-up data (reported in Callaway, 2011). Science retracted a highly publicized field experiment on attitudes to gay marriage when concerns were raised about the authenticity of the data (McNutt, 2015). Medical trials funded entirely by for-profit sources are more likely to find positive results from new treatment compared to existing care than studies funded by nonprofit sources (Ridker and Torres, 2006). The reproducibility project asked researchers to run new experiments to attempt to replicate the results from studies published in top psychology journals in 2008. Of a 100 original studies, only 35 had statistically significant effects in the replication in the same direction as in the original study and the effect sizes in the replication studies had statistically significantly smaller effect sizes (Open Science Collaboration, 2015).

Nor has economics escaped the spotlight. Brodeur et al. (2016) examine studies from top economic journals published between 2005 and 2011 and finds a bunching of results with a p-value just below 0.05, the traditional standard for statistical significance. In chapter "The Production of Human Capital in Developed Countries: Evidence From 196 Randomized Field Experiments" by Fryer (2017) shows a relationship between the magnitude of estimated effect sizes and sample size in published papers of field experiments in education (a telltale mark of publication bias discussed later). Chang and Li (2015) in their examination of macroeconomic papers are only able to "successfully replicate the qualitative findings from 22 of 67 (33%) papers without contacting authors. Excluding the 6 papers with confidential data and 2 papers that use software we do not possess we replicate 29 of 59 papers (49%) with assistance from the authors."

Finally, there have been high-profile arguments about whether there were mistakes in data or analysis and the extent to which the conclusion of several important economics studies should be revised including (in chronological order) Hoxby (2000) (comment Rothstein, 2004, 2005; and response Hoxby, 2007); Donohue and Levitt (2001) (comments Foote and Goetz, 2008; and response Donohue and Levitt, 2006); Rogoff and Reinhart (2010) (comment Herndon et al., 2014; and response Rogoff, 2013); and Miguel and Kremer (2004) (comment Davey et al., 2015; and response Hicks et al., 2015).

It is worth distinguishing between different concerns. Research results may not be reflective of the underlying true state of the world because of the following:

1. data are made up;
2. there was a mistake in the data collection or data analysis;
3. results are not robust to alternative specifications;
4. the findings hold only in a very specific context and are not general;
5. the intervention is not described in enough detail to make it possible to test whether the results hold in a similar context; or
6. sampling variation means the results were due to chance.

One way to address (1) through (3) is to make the data behind a study publicly available. Other researchers can then check the data for signs that it was made up or

manipulated (e.g., Broockman et al., 2015 whose analysis led to the Science retraction mentioned earlier). They can also check that simple mistakes in analysis were not made and that the result is robust to different specifications. Problem (4 requires findings to be tested in different contexts, while (5) requires details of implementation to be included in supplemental material, as journals usually require that this detail is not included in the main paper. Problem (4) can be reduced by adjusting for multiple hypothesis testing and PAPs (discussed in detail later) and by testing the same intervention more than once. However, none of these approaches designed to increase reliability are costless or unproblematic.

In parallel with the concerns about the reliability of original studies, there is also concern about the reliability of attempts to reproduce findings. In commenting on the International Initiative for Impact Evaluation's effort to check the replicability of key international development papers, Ozler (2014) says, "the point of robustness checks in such a replication exercise is not to rerun regressions until you convert one statistically significant result to insignificant and highlight that. … A big part of the point of replication is to reduce p-hacking, not to proliferate it." Commenting on the discussion of the reanalysis of Miguel and Kremer (2004), Blattman in his blog, "Dear Journalists and Policymakers: What you need to know about the Worm Wars," concluded, "Whether it's a sensational photo, a sensational result, or a sensational take down of a seminal paper, everyone has incentives to exaggerate. This whole episode strikes me as a sorry day for science." Simonsohn (2015) points out that many studies that claim to find a "failure to replicate" have a smaller sample size than the original study and are not sufficiently well powered to test whether the original study replicates. Again, there are several different concerns:

1. researchers attempting to test the reliability of an original paper either by reanalyzing the data or running a new study may have incentives to find a result that contradicts the original study;
2. small errors in data or analysis do not always translate into a substantial change in the overall conclusions and it is important to distinguish between meaningful and insubstantial changes in results;
3. if a large number of different specifications are tried in an attempt to test the "robustness" of a result, selective presentation of a few of these specifications may give a misleading impression of how robust the results are;
4. authors may pursue publication only when they find a study does not replicate, giving a misleading impression of the overall reliability of a broad body of research;
5. sample variation may mean the result of a follow-up study is due to chance; and
6. the statistics involved in replication or reproducibility are not straightforward and there is no single agreed standard to judge whether a reanalysis or replication "fails to replicate" the original study.

Confusing the discussion even further is nonstandard use of terminology. Sometimes the term "replication" is used to mean taking the original data and seeing if the same data

generates the tables in the published paper. Sometimes the term is used to mean testing whether the same result is found when the experiment is run on a different sample of the same underlying population. Finally, it could mean testing in a new population. Clemens (2015) provides a useful classification of the different possible options and suggests a standardization of terms. Note that while I try and use Clemens' definitions as much as possible in this chapter, when talking about papers that use different definitions I use the term as used by the author of the paper (especially when quoting papers) (Fig. 1).

4.1 The statistics of data mining, multiple hypothesis testing and publication bias

Before discussing approaches that can be used to address some of the challenges addressed earlier, it is important to be precise about these challenges and the statistics behind them. With the exception of making up data, which is simple fraud, the challenges arise from the fact that standard statistical tests of the significance of the estimated coefficients in a randomized evaluation are based on the assumption that we are testing an independent hypothesis once. In reality, researchers often use one study to test more than one related hypothesis, and one study may not be the only study to test that hypothesis. With full information it is possible to adjust the standard statistical tests to account for the fact that multiple different hypotheses have been tested within a study, or that one hypothesis has been tested multiple times across different studies.

Table 1: A Proposed Standard for Classifying Any Study as a Replication

	Sampling distribution for parameter estimates	Sufficient conditions for discrepancy	Types	Methods in follow-up study versus methods *reported* in original: Same specification	Same population	Same sample	Examples
Replication	*Same*	*Random chance, error, or fraud*	Verification	*Yes*	*Yes*	*Yes*	*Fix faulty measurement, code, dataset*
			Reproduction	*Yes*	*Yes*	*No*	*Remedy sampling error, low power*
Robustness	*Different*	*Sampling distribution has changed*	Reanalysis	*No*	*Yes*	*Yes/No*	*Alter specification, recode variables*
			Extension	*Yes*	*No*	*No*	*Alter place or time; drop outliers*

The "same" specification, population, or sample means the same as *reported* in the original paper, not necessarily what was contained in the code and data used by the original paper. Thus for example if code used in the original paper contains an error such that it does not run exactly the regressions that the original paper said it does, new code that fixes the error is nevertheless using the "same" specifications (as described in the paper).

Figure 1 A proposed standard for classifying any study as a replication.

Most RCTs report both the estimated coefficient on the treatment dummy and the p-value associated with this coefficient. The p-value gives the probability that the estimated coefficient came about by chance. The uncertainty in the estimated coefficient is driven by sampling variation. We randomly sample our treatment and comparison groups from a wider population and we may by chance choose people to include in the treatment group who experience a positive (or negative) shock unrelated to the program we are evaluating. This would lead us to overestimate (or underestimate) the true program effect. If we ran a very large number of RCTs, the average estimated treatment effect would be close to the true treatment effect. An estimated treatment effect from any one trial is one random draw from a distribution of possible treatment effects, centered around the true effect. The probability that any nonzero treatment effect we observe in one particular experiment is due to chance depends on the estimated effect, the sample size, and the variance in the underlying population from which we draw our sample (which we approximate using the sample's variance). The standard calculation for the p-value of an estimated treatment effect assumes that we have made one random draw from the distribution of possible combinations of treatment and comparison groups. If we make more than one draw, we need to be transparent about this and to account for it. (See chapter: The Econometrics of Randomized Experiments by Athey and Imbens (2017) for more discussion on the econometrics behind randomized trials.)

There are two main ways in which our research may deviate from the simple one-arm, one-study assumption behind standard hypothesis testing: a single hypothesis may be tested more than once with several different studies, or multiple different and interrelated hypotheses may be tested in the same study. When we know exactly which hypotheses have been tested by which researchers, it is possible to draw valid conclusions, including by adjusting the calculation of p-values. However, lack of research transparency can lead other researchers to misinterpret the implications of a single study or combination of studies.

4.2 Publication bias

If several RCTs are run on the same population, we are taking multiple draws from the distribution of possible RCTs, and this will increase the precision of our estimate of the true effect size. We will have greater confidence in the weighted average-effect size of all the different studies than in the estimated effect size from one study on its own (where studies with larger sample sizes are given greater weight).[20]

[20] Economists rarely do a formal metaanalysis where coefficients are averaged in this way because we rarely see multiple RCTs of precisely the same intervention on the same population. As Meager (2015) reports, averaging coefficients is not an efficient way to use the data from many studies. Metaanalyses are more common in health, and studies of the same intervention on different populations are averaged based on an assumption that the treatment effect (and underlying population variance) is the same in the different populations. Instead, economists tend to review the studies and discuss how and why treatment effects might or might not vary between populations. Publication bias is as damaging to a metaanalysis as it is to a review of the literature.

However, if we see only a select sample of the RCTs conducted, we may not draw a correct inference about the true effect size. If we see only those realizations that fall in a particular part of the distribution of possible estimated-effect sizes, our overall estimated-effect size will be biased. This selection in the effect sizes we observe can result from researchers seeking to publish only those RCTs that have estimated-effect sizes that fall in a certain range, or if journals only publish those estimated effects that fall in a given range of effect sizes. To illustrate, we take an example where all studies have the same sample size N (and thus should be accorded the same weight) and are done on the same underlying population (and thus are all draws from the same distribution and have the same variance which we assume to be known). Fig. 2A shows the case where the true effect size is zero, and therefore the distribution of possible estimated-effect sizes from RCTs with sample size N is centered around zero. Standard hypothesis testing would give us a critical value $\pm\widehat{\beta}_{cv}$ (i.e., if the estimated effect size is larger/smaller than $\pm\widehat{\beta}_{cv}$, there is less than a 5% probability that the effect size was the result of chance if the true effect was zero). Imagine three different RCTs were run and they provide

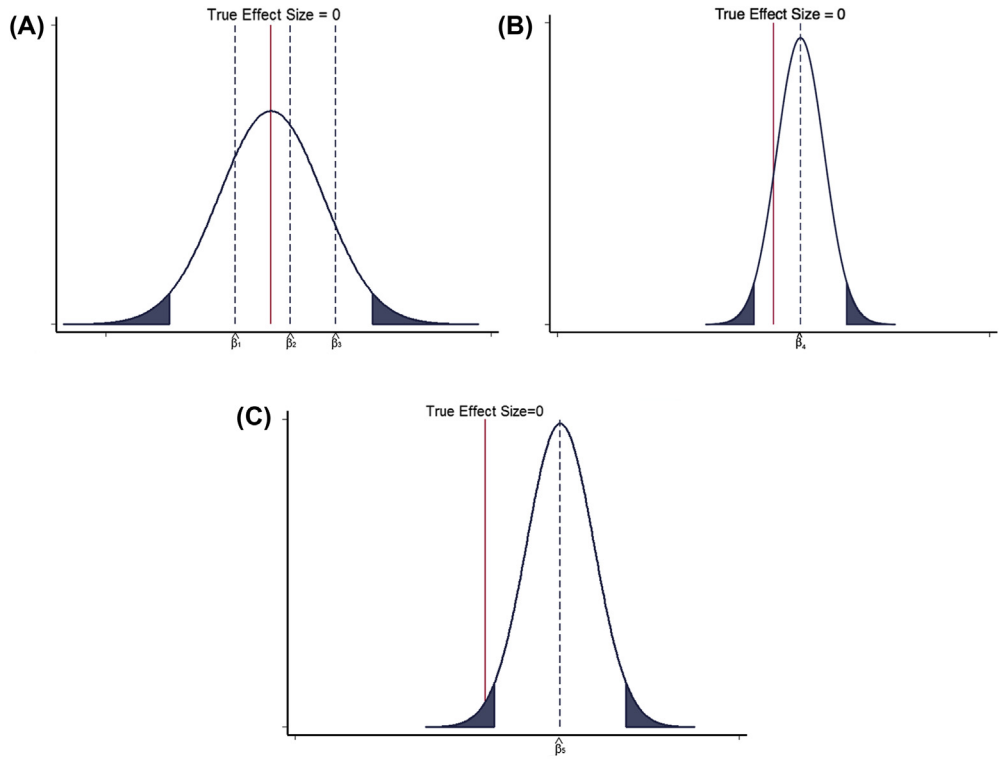

Figure 2 (A) Distribution of possible estimated-effect sizes from RCTs with sample size N centered around zero and a true effect size of zero. (B) Estimated distribution of effect sizes drawn from the results of 3 randomized studies. (C) Estimated distribution of effect sizes drawn from observing only studies 2 and 3.

estimated-effect sizes $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_3$. If we observe all three estimates, we have a new estimated-effect size based on all three studies that has a tighter confidence band than any of the studies on their own. As this confidence band nevertheless includes zero, we correctly fail to reject the null hypothesis that the true effect is zero, and indeed we have reasonable confidence that the true effect is close to zero given our tight confidence band.

However, if those doing or funding the studies have an interest in a certain outcome and repress the results of those studies that do not have a positive coefficient (in this case $\widehat{\beta}_2$) we may reject the null hypothesis, i.e., conclude that the true effect is different from zero with greater than 95% probability. Note that our new estimate of the true effect may be within our original confidence band around zero, but not within our new confidence band, which is smaller given that we are drawing on two studies. Both our estimate of the true effect size and our new confidence band will be biased because of the deliberate exclusion of studies whose estimated effect sizes fall outside a given range. This is one form of publication bias, also known as the "file drawer problem" (Rosenthal, 1979).

Publication bias can also arise if researchers and publishers have no reason to prefer positive or negative results but are more likely to publish results that are significantly different from zero. In the illustration stated previously, only $\widehat{\beta}_3$ is, on its own, significantly different from zero. If only this study is published we might erroneously reject the null hypothesis and conclude the true effect is less than zero. Note that with a large enough set of studies, we will eventually correctly conclude that the estimated effect is indistinguishable from zero even if only studies with results significantly different from zero are published. This is because some studies will be published that have significant positive effects and some will be published with significantly negative effects, and eventually it will be possible to conclude that the correct effect is indistinguishable from zero. However, this process will take much longer than if all studies were published.[21]

Publication bias can be avoided if we know the full population of studies that have been undertaken and know the results of each. This could be achieved by a two-step process in which researchers (1) record the existence of a study and the hypothesis it intends to test before the researcher (or journal) knows the results; and (2) the researcher commits to reporting the results (preferably in the same place) even if the paper never gets accepted in a journal. A researcher doing a literature review could then observe the results of all the RCTs examining a given hypotheses. The second step is harder to police, especially as some journals will not publish a study if the results have been reported elsewhere.

[21] Simply having a longer gap between RCT completion and publication for studies that have results that are, on their own, not significantly different from zero at the traditional confidence levels is sufficient to cause some bias. If there is a stream of RCTs being conducted and a shorter gap between completion and publication for studies that have estimated effects within a given range than for those with estimated effects outside this range, then at any given time a review of published studies will observe a biased set of results and draw inaccurate conclusions.

Thus a researcher may spend years attempting to get a study with a zero treatment effect published and not be able to release the results during that process. Fortunately, even step 1 moves us closer to the goal of reducing publication bias by allowing a researcher undertaking a literature review to observe how many of the studies that sought to test a given hypothesis have had their results published relatively soon after the predicted end of the study. If they observe that all the published studies have positive estimated treatment effects, but that only a small proportion of those that were due to have been completed at the time of the review have been completed, this would cast some doubt on the reliability of the estimated effect of the published studies.

4.3 Current moves to address publication bias

A system of approved registries in which researchers record RCTs that involve health outcomes has been in place for many years: commonly used registries include ClinicalTrials.gov and the EU Clinical Trials Register. An international system for uniquely numbering trials, the International Standard Randomized Clinical Trial Number (ISRCTN), attempts to make it easier to track the number of unique clinical trials on a given topic: trials may have different names and be registered in different places, but they can only have one unique ISRCTN.

The American Economic Association (AEA) recently established a registry for randomized trials in the social sciences (socialscienceregistry.org). The International Initiative for Impact Evaluations (3ie) also has a Registry for International Development Impact Evaluations, which accepts evaluations that are not randomized, but does not accept evaluations of programs in advanced economies. The objective of these registries is to make it easier to track how many studies have attempted to test a given hypothesis in the social sciences. Unlike health journals, social science journals do not (yet) require authors to register their field experiments in an approved registry to be published. However, the AEA and other professional bodies strongly encourage their members to register their trials and a number of funders are now requiring their grantees to do so.

Registering a field experiment is relatively straightforward. The required fields in the AEA registry include title, country, status, trial start and end date, intervention start and end date, a brief description of the experimental design (i.e., the hypothesis to be tested), the main outcomes to be measured, keywords (to allow those doing a literature review to search for all studies that examine a given issue), whether the RCT is clustered, and if so, the number of clusters, the number of planned observations, and whether and from whom human subjects approval was obtained. All of these pieces of information are usually required to obtain human subjects approval to proceed with a field experiment, so the additional burden on researchers of registering is minimal. The registry allows researchers to report the final results on the registry or link to a final paper so that those doing a review can tell whether the results of the study were ever released and what they were.

There is no requirement to provide details on how the data will be analyzed (although it is possible to use the AEA registry to register such a plan, as discussed in the next section). Nor does registering a study mean the authors have to publicly release their data, although the AEA registry does allow for links to published data and the final paper. While it is possible to change information in the AEA registry once a trial has been registered (for example, changing the end date because of delays in program implementation), these changes are tracked so that it is possible to see the evolution of the trial over time. For example, if the sample size is changed, this can be tracked.

If registration is to help mitigate publication bias, it should be completed before the results are known. The AEA labels studies as being "preregistered" if the registration is completed before the intervention starts.

There are relatively few downsides to registration as a way to reduce publication bias. The cost is mainly the time taken to fill out the registration. The main risk is that a registration system may not be useful because it is not complete. While registries for medical studies have operated for a longer time than those in social science and medical journals provide strong incentives for registration, a very large number of studies in these registries never disclose results, even in the registry.

4.4 Data mining and correcting for multiple hypothesis testing

When two different studies test the same hypothesis, it is clear that these represent two draws from the set of possible results, but even within a single study it is possible to effectively take more than one draw.

Imagine we are running a field experiment to test the effectiveness of different health messages in encouraging people to purchase soap. Every day we stand at a grocery store and recruit shoppers into the study. Some are randomized to receive one message and others to receive another message, and we observe their purchases as they check out. Each evening we go home and analyze the data from our field experiment. At the end of the first and second day there is no significant difference in purchasing decisions between those randomized to receive different messages. After the end of the third day we see a significant difference and decide we have reached a big enough sample size to show a significant difference and thus stop the experiment and publish the result. While all three days were part of the same experiment, we are falling into exactly the same trap as described in Section 4.2. We randomly chose three different samples to run our experiment: *day 1* data, *day 1 and 2* data, and *day 1, 2, and 3* data; we decided to show the results of only one of these three—because it produced a result we wanted to see. It is quite possible that this result came from chance variation in who was randomized into which group on day 3. If we had continued the experiment for another day, the difference between the two groups might have gone away again. The solution to this "stopping problem" is relatively simple: we need to define our sample size in advance, based on power calculations, and stop when we reach our predetermined sample size.

To be able to credibly show to others that we followed this procedure it is useful to commit publicly, in advance, to the sample size at which we intend to stop the experiment. The AEA registry is one place where such commitments can be archived. It preserves a record of the date on which the commitment was made and of any changes made to the commitment over time with relevant dates.

The decision about when to stop a rolling enrollment field experiment is only one of many potential choices that a researcher makes about how to collect and analyze data. Many of these choices are, in the case of field experiments, made before the researcher knows what implication these choices will have on the final outcome of the analysis. For example, we make the decision about where to do the study, what type of participants to survey, the sample size, what variables to collect, the time frame over which we expect the impact to become apparent, and how to phrase the questions. Critically, who falls into the treatment and who into the control group is not decided by the researcher.

Decisions over which a researcher has discretion during the analysis stage include whether or not to control for independent variables in the estimating regression; whether to drop "outliers" from the analysis sample and which observations count as outliers; which of potentially many outcome variables to consider the most important; whether to define the outcome measure in levels, logs, or changes; whether and how to combine different outcome measures into an aggregate outcome measure; and within which subgroups to test for heterogeneous treatment effects. The risk that different choices on these issues can lead to different conclusions has been accepted for some time (Leamer, 1983). However, it is important not to overstate this risk, which will vary depending on the situation. With a large enough sample size, controlling for independent variables may somewhat increase the precision of the estimated effect, but choosing different variables to add as controls rarely changes the estimated coefficient much. In most cases, results are not sensitive to whether or not outliers are dropped, and reviewers usually request authors to show that results are robust to including or excluding outliers and controls. In some cases there is also not much discretion about what the main outcome variable should be. A program designed to increase school enrollment will have school enrollment as the main outcome; one designed to improve vaccination rates will likely use percent of children fully vaccinated. While there may be slightly different ways of defining even a seemingly simple outcome measure such as vaccination rates (valid measures include number of children with any vaccinations, proportion of children aged 2–5 years fully vaccinated, proportion of children vaccinated on time, etc.), these measures are usually highly correlated with each other and reviewers will often require the author to show that the results hold for valid, alternative ways of defining the outcome.

A more serious risk of data mining arises when researchers have a concept that has a less precise and generally agreed-upon indicator of success as an outcome. Measurement of concepts such as women's empowerment or social capital may require multiple indicators, with no one indicator being obviously superior to another. In Casey et al. (2012) for

example, we collected over 300 indicators to measure the impact of the GoBifo program on social capital. If we were to consider these indicators separately and run a regression of each potential outcome indicator against our treatment dummy, it is likely that by chance we would find a significant relationship between the treatment dummy and one of these indicators. Indeed, we demonstrate that it is possible to cherry-pick individual outcome indicators which (when taken in isolation) suggest the GoBifo program had positive or negative impacts on a particular aspect of social capital. The true effect from a comprehensive examination of outcome indicators suggests a precise zero impact. If we report all 300+ regressions it would be pretty clear that for the vast majority of outcomes the estimated effect size was zero, and that the few that show significant coefficients (some positive and some negative) were probably the result of chance. If, however, we ran estimating regressions for 300+ potential outcome variables and reported only those where the coefficient was positive and significantly different from zero, we could give the impression that the program was effective in changing social capital, when in fact the data do not support this conclusion. Running many regressions and only reporting those that produce a significant coefficient is often called "data mining," "phishing," or "p-hacking."

There are three basic approaches that can be used to avoid data mining when there are multiple potential ways of defining the main outcome of a study. The first is to combine many outcome variables into a few aggregate outcome variables; the second is to adjust p-values for the fact that multiple hypotheses are being tested; and the third is to commit in advance to how the data from an experiment (or other analysis) will be analyzed.

The simplest way to combine many potential outcome variables into one is to create an index. We may collect many indicators designed to measure wealth, including a series of asset dummies that take the value 1 if a household owns a radio, or bike, or TV. Rather than test the impact of a program on each individual asset dummy, we create a wealth index that is the mean of all the individual asset dummies. We then estimate the impact of the program on the overall wealth index. The same can be done for other multifaceted outcome measures. For example, we may ask a series of questions about whether a woman is involved in various household decisions. We can create a decision-making index by averaging the responses to all these questions. Indices are usually used to combine many similar dummy variables.

A mean effects approach used by Kling et al. (2007) in their evaluation of Moving to Opportunity is an alternative and increasingly popular way to combine outcome indicators that are in a similar "family" of outcomes. A family of outcomes may be ones that all ask about health or education or another similarly broad topic. To estimate mean effects, all the variables in a family need to be placed on a similar scale so that each has the same mean (zero), standard deviation, and direction (negative should be bad for all variables, and positive good). We then run a linked set of estimations on the new set of variables, and the "mean effect" is the average of all the coefficients in the set of linked estimations.

An index of mean effects can be used to reduce the number of outcome measures for which we estimate a treatment effect. Having reduced our hypotheses to a manageable number, we can adjust the p-values for the fact that we are testing several related hypotheses. The Bonferroni correction is the simplest way to do this but it suffers from low power: we may fail to reject the null even when we should. A better approach is to use the free step-down resampling method for the family-wise error rate (Westfall and Young, 1993 are credited with the approach, and Anderson, 2008 provides a good explanation of its use). One advantage of this latter approach is that it takes into account that the outcome variables may be correlated with each other.

Adjusting p-values when we present different outcome measures is not always necessary or appropriate. Some hypotheses are clearly secondary and are designed to illustrate the mechanism through which the main effect was achieved (or was not achieved). In Banerjee et al. (2010a,b,c) we evaluated a Pratham program designed to increase reading levels by providing information to parents about the poor reading levels of students and ways in which they could advocate for change. We collected outcomes on whether the information was provided to parents, whether parents changed their beliefs about how much their child was learning, whether parents put in more effort to monitor schools or advocate for more education resources, whether more resources were secured for schools, and whether test scores increased. We created families of outcomes for variables associated with each step in the process, but we did not collapse them into one family, nor did we adjust the p-values for the fact that we had several families. This is because even though we found statistically significant results for the first two outcomes (information sessions happened and parents were more informed), we did not declare the program a success because the causal chain clearly stopped at this point. Being better-informed did not lead to more effort, resources, or outcomes. In other words, results are always judged in the context of theory, and this can be an important barrier to data mining. Note that some authors fail to make this distinction between a paper's main outcome and regressions that test mechanisms. Thus Young (2016) "corrects" the significance for various RCTs by adjusting p-values for all the regressions reported in the paper whether or not these regressions tested a main outcome.

The other key area where data mining is a particular risk (and being accused of data mining is a serious risk) is subgroup testing. As with multiple outcomes, testing for differential effects among subgroups raises concerns about multiple hypotheses. If we simply test for effects in every possible subgroup, we will likely find one with a significant treatment effect. One option is to adjust p-values for the number of subgroups tested. A better approach is to have a clear motivating theory behind why some subgroups will react differently than others. For example, if we are evaluating a program that provides incentives for girls to stay in school and we find that the program had a bigger effect on girls who are within walking distance of a secondary school than those who are not, this will strengthen our confidence in the overall result.

4.5 Preanalysis plans

Perhaps the most robust way to avoid data mining, or being accused of data mining, is to commit in advance to how the data from a field experiment will be analyzed by creating a PAP. A PAP can be a useful complement to the strategies discussed earlier. For example, if we plan to create five families from 300+ outcome variables, we have a large amount of discretion about how to divide them up unless we commit in advance which variables will go into which families. It is hard to credibly adjust our p-values for the number of regressions run unless we commit in advance to exactly which regressions we intend to run (without this we could run more and only pick the ones that were significant, and then adjusting our p-values for those would be meaningless). Similarly, if we want to adjust our p-values for the number of subgroup analyses we run, it is important to state at the start which subgroups we intend to test.

There has been an increase in the use of PAPs among those doing field experiments, but they are far from the norm. Many economists feel PAPs are too constraining, that authors discover important truths in the data that they could not have predicted prior to their examination of it, and that it would be wrong not to pursue these revelations. Others worry that "following the data" in this way can lead researchers to find patterns that are there just by chance and that tying their hands in advance is useful.

Writing a PAP is not without costs. It is a time-consuming and difficult process. It is hard to think through what additional tests should be carried out under each combination of possible results, especially when a trial has multiple arms. As Olken (2015) explains, "Most research papers test a large number of hypotheses. Hypotheses themselves are often conditional on the realization of other, previous hypothesis tests: the precise statistical question a paper might test in Table 4 depends on the answer that was found in Table 3; the question posed in Table 5 might depend on the answer in Table 4, and so on."

There is nothing to stop a researcher from including results to questions that were not posed in the PAP in a paper. These can be considered exploratory rather than confirmatory. However, most readers will put less weight on these results than they would put on a result in a paper without a PAP. While the abstract scientific model would suggest that subsequent research can follow-up such exploratory results with confirmatory tests, Olken notes that such follow-up work is less common in economics than in medicine, not least because funding for economic research is a fraction of that in medicine. Olken also argues that there is an opportunity cost to researchers thinking through how they would analyze results under multiple different scenarios ahead of time and that this may come at the cost of focusing more deeply on the one scenario that is revealed once the data are analyzed. This may be less a question of the opportunity cost of time than the risk that the PAP process makes a researcher fix ideas ahead of time and is thus less flexible to seeing patterns they had not thought of prior to exploring the data. Indeed, this "fixing ideas ahead of time" is precisely the benefit of a PAP (it reduces

the risk of being persuaded by patterns that are there by chance) but also the cost (maybe the patterns are not there by chance, but we miss them if we are blinded by our PAP). This is the fundamental trade-off in PAPs.[22]

Given these trade-offs, the most common use of PAPs is for field experiments in which there is no obvious single, primary outcome variable, or where the authors know that subgroup analysis will be a critical part of their paper and are nervous of being accused of data mining. Given the complexity of doing PAPs in multiarm studies, PAPs are also more common in one-arm trials. PAPs are also used by researchers to help manage relationships with partners. It can be very helpful to have a written document that clarifies what outcome the partner hopes or expects to see and would count as success. This can prevent awkward discussions later in which the partner wants to cherry-pick positive findings. These partner/researcher documents do not necessarily have to be as detailed as a full PAP to be useful (an example of a broad partner/researcher agreement followed by a detailed PAP can be found in Casey et al., 2012).

Olken (2015) provides a useful checklist for what should be included in a PAP:

Preanalysis plan checklist

Item	Brief description
Primary outcome variable	The key variable of interest for the study. If multiple variables are to be examined, one should know how the multiple hypothesis testing will be done.
Secondary outcome variable(s)	Additional variables of interest to be examined.
Variable definitions	Precise variable definitions that specify how the raw data will be transformed into the actual variables to be used for analysis.
Inclusion/Exclusion rules	Rules for including or excluding observations and procedures for dealing with missing data.
Statistical model specification	Specification of the precise statistical to be used, hyperthesis tests to be run.
Covariates	List of any covariates to be included in analysis.
Subgroup analysis	Description of any heterogeneity analysis to be performed on the data.
Other issues	Other issues include data monitoring plans, stopping rules, and interim looks at the data.

Olken, B.A., 2015. Promises and perils of pre-analysis plans. J. Econ. Perspect. 29 (3), 61–80.

[22] A related but different issue is that if all the regressions committed to in the PAP are reported in the main paper, this can make for a boring read. Imagine, for example, that treatments 1, 2, and 3 are different twists on a base program and the treatment effects on all three are all insignificant from zero, as is the coefficient on the pooled outcome and all subgroups. We really do not need to see all of these results; we just need to see the result from the three-pooled arms (which has the most precise estimate) and a footnote saying none are significant when run separately, and none of the subgroups are significant. Some authors put all the results specified in the PAP in an appendix but do not necessarily show them all in the main text to help with this problem.

One issue that is still debated within economics is the best time during the research process to write a PAP. A purist approach would suggest that the PAP should be written before the start of the experiment, but it is not clear that this is optimal. Casey et al., for example, argue that there are several advantages to waiting: the literature may have advanced during the trial which may raise additional hypotheses that can be tested with the data generated in the trial; observations on the ground may also generate additional hypotheses that can be tested, including unforeseen negative impacts of the intervention; and the process of baseline data collection can also inform the researcher about which outcome variables are well measured and for which outcome measures there is room for improvement.

In FDA-regulated trials, only the primary and secondary outcome variables are specified prior to the start of the trial, while the detailed data handling and analysis plan is written after the end line data are collected, but before the data are combined with information on which observations are treatment and which are comparison (Olken, 2015). This allows the researcher to determine the best-fit specification prior to including treatment status, or drop outliers before the researcher knows whether the outliers are treatment or control. Some economists have used this approach (Olken in particular recommends it), while others prefer to set the PAP before the end line is collected. Bidwell et al., (2015) use a multistage PAP for a paper that included several rounds of data collection: an initial overarching PAP was written and was then updated at prespecified times after the analysis of a given data set raised hypotheses which could then be tested in subsequent data sets.

A number of PAPs for field experiments are now publicly available and are worth examining before writing one for the first time. Some of the early PAPs in economics can be found at http://www.povertyactionlab.org/Hypothesis-Registry. These include PAPs for Targeting the Poor (published as Alatas et al., 2013), GoBifo (published as Casey et al., 2012), and the Oregon Health Insurance Experiment (published as Finkelstein et al., 2012). Since the opening of the AEA Registry, new PAPs in economics have been published at https://www.socialscienceregistry.org/.

4.6 Evidence on the magnitude of the problem

Increasing the transparency and reproducibility of research results is not costless. Preparing data for publication takes time that could be devoted to doing new studies. The same is true of reanalyzing data and running reproduction and extension studies. Whether these costs are worth incurring depends in part on how big the problems are compared to the costs. Estimating the magnitude of the problem is not easy. We cannot judge the magnitude of the problem by examining published studies of replication or reanalyses and asking how many of these published articles find the study replicates and how many claim

to have found a failure to replicate. Replication or reproducibility efforts can be subject to the same issues of publication bias and data mining. If anything, the incentives for publication bias and data mining may be worse for reproduction studies than for original studies. A zero effect in an original study may not be as exciting as a large positive or negative impact, but it is at least a new finding. If a replication study finds exactly the same effect as the original study it does not even have the benefit of being news and an author may well not put a lot of effort into trying to get it published, or worse may attempt to manipulate the results to show that the original finding is not robust. In other cases, replication studies have been much less well powered than the original study. Failing to find a significant effect in a low-powered study when the original study found a significant effect is not a "failure to replicate" as is too often claimed. Simonsohn (2015) also points out that testing whether the two estimated-effect sizes in the different papers are significantly different from each other may also not be a good way to judge if the new study fails to replicate the first study. He suggests that the appropriate standard is whether the replication results are consistent, i.e., that there is an effect size that is large enough to be detectable by the original study. Simonsohn argues that much of the evidence for bias in the psychology literature is based on inappropriate tests. For example, all 10 of the most cited studies in psychology that use "failure to replicate" in their title use as their test whether the replication study is significantly different from zero even though the replication studies often have substantially smaller samples than the original study.

A way round this publication bias in replications is to define a specific set of studies that are to be reanalyzed or reproduced and set clear standards in advance about how the reanalysis or reproducibility is to be judged. An additional benefit of this approach is that there is little incentive among the reproducers to either confirm or undermine the initial result: the result will surprise some people whether the finding is that many studies can be reproduced or few can be reproduced.

Two large initiatives to reproduce findings from psychology studies have recently concluded. Klein et al. (2014) had multiple labs retest important findings in psychology. Some of the labs were in the same country as the original experiment (i.e., were reproductions, in Clemens' parlance) and some were on new populations (extensions, according to Clemens' definition). Of the 13 original findings, similar results were found for 10, reasonably similar results were found for one, and in 2 cases there was little evidence of a consistent result holding either in the original or new population. The effect sizes in the replications were sometimes larger and sometimes smaller than the original. Combined, these results are rather encouraging.

A second initiative in psychology looked at a larger number of studies (100) but attempted to reproduce them only once and mostly on similar populations (Open Science Collaboration, 2015). With one original study and one reproduction where

the results are not consistent, it is not possible to say which is the correct result. However, it is possible to examine patterns across the many pairs of original and reproduction studies. As shown in the figure of a study by Open Science Collaboration (2015) it is possible for a replication to find an effect size of similar magnitude to the original study but not be significantly different from zero. It is also possible for the replication to generate an effect size that is significantly different from zero but not be of similar magnitude. Neither is a good measure of reproducibility on its own. However, what the figure in the study by Open Science Collaboration (2015) also shows is that, on average, replication studies had substantially lower estimated effect sizes than original studies and had substantially lower significance. This is true even if we look just at the studies with high power. Combined, these results raise important concerns about reproducibility. What explains the difference between the findings of these two initiatives is unclear. One possible theory is that Klein et al. looked at more famous results, and that these results are famous for a reason. This contention is supported by Dreber et al. (2015), which shows psychology researchers are able to predict reasonably well which of the 100 studies would replicate.

No such well-structured, systematic attempt to reproduce economics studies has been undertaken. Instead, Brodeur et al. (2016) use a different approach to estimating the bias in field experiments in economics. If researchers' data mine to tip their results just above a critical significance level of 5% or 10%, or if studies with results under these levels are less likely to be published, there will be few published results just below these cutoffs. By examining empirical studies from three top journals (*AER*, *JEP*, and *QJE*) between 2005 and 2011, Brodeur et al. find evidence of this "missing mass" just below conventional significance levels in nonrandomized studies in economics, but not in field experiments. Olken (2015) argues that even the level of manipulation observed in nonrandomized studies in economics suggested by the missing mass is not substantial. Brodeur et al. estimate that 10–20% of p-values below 0.05 should in fact be between 0.10 and 0.25. Olken points out that this means that of a 100 studies, instead of having 5 false rejections, we would have 7.25 false rejections. He argues that while this is not an ideal outcome, it suggests that actions to address publication bias and data mining should only be taken if they do not impose a large cost on research.

Why do we not see more evidence of p-hacking? One explanation is that the long process economics papers go through prior to publication helps reduce the ability to p-hack. Papers are typically presented many times with robust discussion and criticism. During this and the referee process, if an obvious specification is not included in robustness checks, seminar participants and referees will usually ask to see it. Authors are not just motivated by how many publications they have but by the respect of their peers, and any appearance of twisting the data to get a certain result loses an author the respect of their

peers. Now that data are more commonly published alongside the article, authors may be conscious that others will try different robustness checks, and that they will be criticized if the results are not found to be robust. Finally, financial stakes are not as high in economics as they are in medicine, where multimillion dollar revenue streams rest on a drug or device being found effective. This does not mean there is no p-hacking in economics or that the referee system works perfectly; only that we need to keep the magnitude of the problem in perspective.

While Brodeur et al.'s results and similar analyses of the distribution of p-values can tell us about p-hacking as well as possible publication bias around specific cutoffs like 0.05, they are not very informative of the overall magnitude of publication bias. While Brodeur et al. find fewer studies with p-value of 80% than 4%, this does not necessarily signify publication bias; it may signify that authors are more likely to test for plausible relationships than implausible relationships. If we want to test the magnitude of publication bias, we need to find a defined sample of studies of similar quality when they are started and follow up which ones make it to publication.

Franco et al. (2014) do exactly this: they follow up all 221 research proposals that won a competitive award to get access to a representative sample of the US population with the objective of running an experiment under the Time-sharing Experiments in the Social Sciences initiative. Of 49 studies that produced null results, only 10 were published in journals and 1 as a book chapter. Of 93 studies with strong results, 56 were published as journal articles and 1 as a book chapter. Much of the difference between the two groups can be explained by whether the authors wrote up the results. For example, only 7 of the 38 unpublished studies with null results were even written up. The last finding may simply be authors internalizing that journals are unlikely to publish the study if it was written up. It may also reflect some unobserved heterogeneity in quality of study, although the authors argue this is less of a concern given the tough competition to get these grants.

It is also the case that some findings are less interesting than others, and researchers and journals should place more emphasis on important and interesting results. An out-of-the-box idea may be interesting if it has a significant impact effect, but not if it has zero effect. This opportunity cost of time and journal space has to be set against the costs of publication bias.

4.7 Incentives for replication and transparency

In this section we discuss the extent to which there are sufficient incentives for researchers undertaking field experiments in the social sciences to be transparent, and what, if any, additional incentives should be put in place to encourage transparency, reanalysis, robustness testing, reproduction, and extension work.

Of all the strategies for creating greater transparency, registration of field experiments is probably the least costly: it takes little time and is unlikely to distort research.[23] Because it is not hard, it may not require large incentives. However, as the benefits are public, a nudge is appropriate, especially to get experiments registered early. Increasingly, funders and research-implementing organizations are requiring registration.

As discussed earlier, PAPs are hard and costly, and there is little evidence of statistical data mining in field experiments in economics. The main incentive to do a PAP is for the author to protect themselves from accusations of data mining. Further incentives at this stage are unlikely to be warranted.

The benefits of data publication are potentially larger, as it allows for checks of robustness and fraud, and allows others to do research on related issues. The costs to researchers are reasonably high, but are one-time costs. Publication does not distort research inappropriately. It is here that incentives should be focused.

Twenty or even ten years ago there was little expectation among economists that the data behind a paper would be made available at the time of publication. There were exceptions. For example, the MacArthur Foundation Network on Inequality in the late 1990s funded a series of studies in development, including some of the early field experiments, and required that data from all the studies were published (Research Network on Economic Equality & Social Interactions). It was also the case that many economists in development felt obliged to make their data available when others requested it.

The incentives to make data publicly available have increased rapidly. In 2004, the *American Economic Review* started requiring authors to make public the data and replication code in support of the tables in the published paper. Most top journals followed, as well as applied field journals such as the *Journal of Labor Economics* and the *Journal of Development Economics*. Knowing data publication will be required if the paper does well can encourage good data management and documentation along the way, which then makes the task of publication easier. Many funders, including the International Initiative on Impact Evaluation, the Arnold Foundation, and the Abdul Latif Jameel Poverty Action Lab, require studies they fund to publish their data.

Exactly what counts as data publication is still debated. Most economics journals only require authors to post the part of the data set that is necessary to replicate the tables in the paper. Nor do they require the original raw data to be posted. This means that some robustness checks cannot be carried out (for example, whether the result holds when controlling for a variable that is collected but not included in the controls and thus not published). It also means that some manipulation can happen during the creation

[23] Coffman and Niederle (2015) suggest that a possible cost of registries is that authors may not want to share their research designs before the paper is published. However, the AEA registry allows authors to hide details of design until the paper is published.

of aggregate variables or in "cleaning" the data. But it is not clear that posting raw data is in fact more transparent. Most raw data requires so much work that it is impenetrable to anyone not involved in the study. Even if the raw data and the code to turn it into clean data are posted, the cleaning files will be so long and tedious that it is unlikely anyone will learn anything useful from them. As an example, in a nationally representative survey of smallholder agriculture in Sierra Leone, cleaning involved thousands of manual corrects for double-entry reconciliation errors, as well as turning a dozen different local ways of measuring output into a common standard. It would be hard for another researcher to comment on how much bigger a Kenema buttercup is than a Bo buttercup and how many there are in a bushel. In psychology, authors have pointed out that some raw data consist of brain scans and posting all the brain scans in a study would be infeasible.

More concerning is the finding from Chang and Li (2015) detailed earlier that many studies published in journals with data publication requirements do not have sufficient information posted to allow authors to reproduce the results in the paper. In addition, many data sets sit unpublished for lack of a few days additional work, even though this is a tiny proportion of the work that went into collecting, cleaning, and analyzing data. These data represent an important public good: in addition to being useful in checking the validity of published findings, they can be used by other researchers to calculate intracluster correlations for power calculations, combine with other data to do new analysis, publish descriptive studies, or explore relationships not explored by the original authors. An effort is now underway by various institutions, including the Berkeley Initiative for Transparency in Social Sciences, Innovations for Poverty Action, and the Abdul Latif Jameel Poverty Action Lab, to provide assistance and financial incentives to get more data published. Papers such as Chang and Li, which audit journals' implementation of publication rules, are also useful.

Should anything be done to incentivize the researchers to use data that are published to check the validity of existing studies? One position is that academic incentives are skewed toward producing new studies and thus there is not enough checking of existing studies. This was one reason that 3ie launched their project to fund researchers to attempt to verify existing studies. As we discussed earlier, however, many criticized this attempt: the incentives were for those doing the verification and robustness checks to find problems and the results were published on the 3ie Website without going through peer review; verification attempts that found no problems appear to be less likely to be published; and there were accusations that verification authors departed from their PAPs. (In the interests of full disclosure I am married to Michael Kremer whose paper was part of this process and the comment on which started the worm wars.)

PhD students undertaking verification as part of their studies is a process widely regarded as less subject to incentive problems. Students learn a lot from the process and will do as well in the class if they achieve verification as if they find an error. The

only disadvantage is that there is no public record of what papers are verified in this way: The profession has a sense that many papers are tested and only a few headline errors have come out, but no one knows for sure how many have been tested. This lack of a record also means that different classes may be verifying the same paper again and again while others go unverified. Professors teaching these classes have expressed concern in going public with the results of all their students' work, explaining that students do not always have the time or the ability to do a thorough verification job. Nevertheless, a simple record of which papers have been the subject of scrutiny in a given class, without indicating the results, could be a useful step toward transparency. If a paper has been examined in multiple classes without any published comment about potential errors, this could help increase confidence in the result even if no one student paper should be taken as meaning a clean bill of health.

An even lower pressure, incentive-aligned strategy is to have the verification take place before publication. Authors could submit data and code to an independent team that attempts to verify the analysis prior to publication. This could be done through journals themselves or through independent groups. Again, PhD students or postdocs might be happy to do this work if it was as well paid as teaching assistant jobs providing them with the income they need to pursue their own research. JPAL is now undertaking prepublication verification of selected papers.

This leaves the more complex and expensive problem of encouraging reproductions or extensions, i.e., running a new experiment in either the same/similar or a different population. Coffman and Niederle (2015) conclude that reproductions are particularly helpful. They address issues of p-hacking, errors in data and analysis, and the risk that a result is due to chance. They consider them less distortionary to the research process than PAPs, which they argue have considerable costs in preventing flexibility in analysis. They recognize, however, that large field experiments may be too expensive and hard to undertake multiple times. Additionally, academics may not have the incentive to do the second, third, or fourth study on a given issue. One approach is to fund well-coordinated efforts that test an approach in many different contexts at once. The resulting collection of results may well attract significant attention and academic reward. For example, the coordinated series of studies on the graduation approach pioneer by BRAC was published in *Science* (Banerjee et al., 2015b). These coordinated approaches are expensive and hard. They face the constant tension between the goal of testing a very standardized program across contexts and allowing the program to be adapted to local needs and preferences. Coffman and Niederle (2015) propose having a new journal that only publishes reproductions and possibly extensions. This has the advantage that the new studies are peer reviewed for quality and authors know there is an academic outlet for their reproduction studies.

A particularly interesting part of their suggestion is that any subsequent citation of the original study could be followed with notation indicating whether the result has been reproduced (R+ for one that has been successfully reproduced; R− for one that has been unsuccessfully reproduced). This means there are some upsides for the original author in having their study reproduced. They admit there are many issues to sort out, such as what counts as a successful reproduction. They also note that reproductions are likely to be more common for studies that are cheap to run. We would add this approach is more useful when the intervention is very clearly defined, as this is the only way to ensure that exactly the same intervention is being tested in the original and the reproduction. Very large, complex, and expensive studies may never, or rarely, have an attempted reproduction. Another, decentralized approach, which is currently more typical in economics, is for different researchers to test the predictions of a single theory in different ways in different contexts. The theory-driven approach is less useful for testing programs with complex interdependent components, as in the graduation program, where multisite studies may be more useful. The decentralized theory-driven testing is more likely to remain more prevalent, and academics continue to have strong incentives to test theories posed in one paper in similar and different contexts.

The debate about increased transparency and reproducibility in economics too often fails to apply this more theoretical lens, and in doing so can give the impression that we know less than we do. For example, we may have few exact replicas testing whether the incentives for immunization program in India tested by Banerjee et al. (2010c) "works" in other countries. But we do have multiple studies from different countries testing the same underlying hypothesis that small changes in price (both positive and negative) can have surprisingly large impacts on the take-up of health prevention products (for a summary, see Kremer and Holla, 2008; Kremer and Glennerster, 2011). Similarly, the recent series of studies on the impact of providing voters information about candidates prior to elections all test whether voting is purely clientalistic (Fujiwara and Wantchekon, 2013; Bidwell et al., 2015; Ferraz and Finan, 2007; Banerjee et al., 2010a,b,c). We would not want to provide exactly the same information to voters in the same way in different countries. Nor would we want to test whether the coefficient found in later ("replication") studies were significantly lower than those in earlier ("original") studies, or test if one study is within the margin of error of that found in another study. However, the fact that studies in different developing countries have consistently found information provision changes how people vote provides us with more confidence in the reliability of the finding than one study on its own.

The classic approach in economics is, instead of testing whether a program "works" across contexts, to test whether theories hold in a variety of situations.

5. CONCLUSION

Field experiments are hard to do well, and the majority of blood, sweat, and tears come in the details of research implementation. Poor judgment during any one of the thousands of small decisions can undermine the entire venture. Attention to detail is critical. A miscommunication with a partner can lead to the randomization protocol not being followed, an underestimated budget will mean the project cannot be completed as envisaged, a badly worded survey question can lead to an ambiguous outcome measure, low take-up can cause the experiment to be underpowered, or a survey conducted at the wrong time of year can lead to high attrition rate. The role of the researcher is not just to design the evaluation, but to be on top of these practical decisions throughout the process, from design to data publication.

REFERENCES

Alatas, V., Banerjee, A., Hanna, R., Olken, B.A., Purnamasari, R., Wai-Poi, M., 2013. Ordeal mechanisms in targeting: theory and evidence from a field experiment in Indonesia. Natl. Bur. Econ. Res.

Alderman, H., Das, J., Rao, V., 2013. Conducting Ethical Economic Research: Complications from the Field. World Bank Policy Research Working Paper, No. 6446.

Aldermanet, H., Das, J., Rao, V., 2014. Conducting Ethical Economic Research: Complications from the Field. The Oxford Handbook of Professional Economic Ethics, Oxford, UK. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199766635.001.0001/oxfordhb-9780199766635-e-018.

Allcott, H., 2015. Site selection bias in program evaluation. Q. J. Econ. 130 (3), 1117–1165. http://dx.doi.org/10.1093/qje/qjv015.

Anderson, M.L., 2008. Multiple inference and gender differences in the effects of early intervention: a reevaluation of the abecedarian, Perry preschool, and early training projects. J. Am. Stat. Assoc. 103 (484).

Angrist, J., Bettinger, E., Kremer, M., 2006. Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia. Am. Econ. Rev. 847–862.

Angrist, J.D., 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. Am. Econ. Rev. 313–336.

Ashraf, N., Berry, J., Shapiro, J.M., 2010. Can higher prices stimulate product use? Evidence from a field experiment in Zambia. Am. Econ. Rev. 100 (5), 2383–2413. http://dx.doi.org/10.1257/aer.100.5.2383.

Athey, S., Imbens, G.W., 2017. The econometrics of randomized experiments. In: Duflo, E., Banerjee, A. (Eds.), Handbook of Field Experiments, vol. 1, pp. 73–140.

Baird, S., Hamory, J., Miguel, E., 2008. Tracking, attrition and data quality in the kenyan life panel survey round 1 (KLPS-1). Cent. Int. Dev. Econ. Res.

Banerjee, A., Chattopadhyay, R., Duflo, E., Keniston, D., Singh, N., 2012. Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy and Training. w17912. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w17912.pdf.

Banerjee, A., Duflo, E., Glennerster, R., Kinnan, C., 2015a. The miracle of microfinance? Evidence from a randomized evaluation. Am. Econ. J. Appl. Econ. 7 (1), 22–53. http://dx.doi.org/10.1257/app.20130533.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., Udry, C., 2015b. A multifaceted program causes lasting progress for the very poor: evidence from six countries. Science 348 (6236), 1260799.

Banerjee, A., Hanna, R., Kyle, J.C., Olken, B.A., Sumarto, S., 2015. Contracting Out the Last-Mile of Service Delivery: Subsidized Food Distribution in Indonesia, w218372015. National Bureau of Economic

Research, Cambridge, MA. https://www.povertyactionlab.org/sites/default/files/publications/553%20Raskin%20Contracting%20Last%20Mile%20NBER%20Dec2015.pdf.
Banerjee, A., Kumar, S., Pande, R., Su, F., 2010a. Do Informed Voters Make Better Choices? Experimental Evidence from Urban India. Unpublished Manuscript. http://www.Povertyactionlab.org/node/2764.
Banerjee, A.V., Banerji, R., Duflo, E., Glennerster, R., Khemani, S., 2010b. Pitfalls of participatory programs: evidence from a randomized evaluation in education in India. Am. Econ. J. Econ. Policy 1–30.
Banerjee, A.V., Duflo, E., Glennerster, R., Kothari, D., May 17, 2010c. Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. BMJ 340 (1), c2220. http://dx.doi.org/10.1136/bmj.c2220.
Barrett, C.B., Carter, M.R., 2014. Retreat from radical skepticism: rebalancing theory, observational data and randomization in development economics. In: Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences, pp. 58–77.
Beaman, L., Keleher, N., Magruder, J., 2013. Do Job Networks Disadvantage Women? Evidence from a Recruitment Experiment in Malawi. Working Paper. Department of Economics, Northwestern University.
Beath, A., Christia, F., Enikolopov, R., 2013. Winning hearts and minds through development: evidence from a field experiment in Afghanistan.
Bidwell, K., Casey, K., Glennerster, R., June 2015. Debates: The Impact of Voter Knowledge Initiatives in Sierra Leone. Abdul Latif Jameel Poverty Action Lab Working Paper. http://www.povertyactionlab.org/publication/debates-impact-voter-knowledge-initiatives-sierra-leone.
Blattman, C., October 23, 2015. Dear Journalists and Policymakers: What You Need to Know about the Worm Wars. Chris Blattman Blog. http://chrisblattman.com/2015/07/23/dear-journalists-and-policymakers-what-you-need-to-know-about-the-worm-wars/.
Board of Governors of the Federal Reserve System, Chang, A.C., Li, P., 2015. Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'. Finance and Economics Discussion Series 2015 (83), pp. 1–26. http://dx.doi.org/10.17016/FEDS.2015.083.
Brodeur, A., Lé, M., Sangnier, M., Zylberberg, Y., 2016. Star wars: the empirics strike back. Am. Econ. J. Appl. Econ. 8 (1), 1–32. http://dx.doi.org/10.1257/app.20150044.
Broockman, D., Kalla, J., Aronow, P., 2015. Irregularities in LaCour. https://web.stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf.
Bruhn, M., McKenzie, D., 2009. In pursuit of balance: randomization in practice in development field experiments. Am. Econ. J. Appl. Econ. 1 (4), 200–232. http://dx.doi.org/10.1257/app.1.4.200.
Buchmann, N., Field, E., Glennerster, R., Nazneen, S., Pimkina, S., Sen, I., 2016. The effect of conditional incentives and a girls' empowerment curriculum on adolescent marriage, childbearing and education in rural Bangladesh: a community clustered randomized controlled trial. Abdul Latif Jameel Poverty Action Lab Working Paper December 2016. https://www.povertyactionlab.org/sites/default/files/KK_empowerment_Bangladesh_Dec2016%20%281%29.pdf.
Callaway, E., 2011. Report finds massive fraud at Dutch universities. Nature 479 (7371), 15. http://dx.doi.org/10.1038/479015a.
Casey, K., Glennerster, R., Miguel, E., 2012. Reshaping institutions: evidence on aid impacts using a pre-analysis plan. Q. J. Econ. 127 (4), 1755–1812. http://dx.doi.org/10.1093/qje/qje027.
Chandrasekhar, A., Kinnan, C., Larreguy, H., 2015. Social Networks as Contract Enforcement: Evidence from a Lab Experiment in the Field. Working Paper. http://faculty.wcas.northwestern.edu/~cgk281/SaI.pdf.
Chattopadhyay, R., Duflo, E., 2004. Women as policy makers: evidence from a randomized policy experiment in India. Econometrica 72 (5), 1409–1443.
Clemens, M.A., 2015. The Meaning of Failed Replications: a Review and Proposal. Institute for the Study of Labor (IZA).
Coffman, L.C., Niederle, M., 2015. Pre-analysis plans have limited upside, especially where replications are feasible. J. Econ. Perspect. 29 (3), 81–98. http://dx.doi.org/10.1257/jep.29.3.81.
Cohen, J., Dupas, P., 2010. Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. Q. J. Econ. 125 (1), 1–45. http://dx.doi.org/10.1162/qjec.2010.125.1.1.

Cole, S.A., Fernando, A.N., 2012. The value of advice: evidence from mobile phone-based agricultural extension. SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.2179008.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., Zamora, P., 2012. Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment. w18597. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w18597.pdf.

Davey, C., Aiken, A.M., Hayes, R.J., Hargreaves, J.R., July 2015. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. Int. J. Epidemiol. http://dx.doi.org/10.1093/ije/dyv128 pii:dyv128.

Dhaliwal, I., Hanna, R., 2014. Deal with the devil: the successes and limitations of bureaucratic reform in India. Natl. Bur. Econ. Res.

Donohue, J.J., Levitt, S.D., 2001. The impact of legalized abortion on crime. Q. J. Econ. 116 (2), 379–420. http://dx.doi.org/10.1162/00335530151144050.

Donohue, J., Levitt, S., 2006. Measurement Error, Legalized Abortion, and the Decline in Crime: a Response to Foote and Goetz (2005). w11987. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w11987.pdf.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B.A., Johannesson, M., November 2015. Using prediction markets to estimate the reproducibility of scientific research. Proc. Natl. Acad. Sci. http://dx.doi.org/10.1073/pnas.1516179112.

Duflo, E., Gale, W., Liebman, J., Orszag, P., Saez, E., 2005. Saving incentives for low-and middle-income families: evidence from a field experiment with H&R block. Natl. Bur. Econ. Res.

Duflo, E., Saez, E., 2002. Participation and investment decisions in a retirement plan: the influence of colleagues' choices. J. Public Econ. 85 (1), 121–148. http://dx.doi.org/10.1016/S0047-2727(01)00098-6.

Duflo, E., Greenstone, M., Pande, R., Ryan, N., 2013. Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India, w192592013. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w19259.pdf.

Fearon, J.D., Humphreys, M., Weinstein, J.M., 2009. Can development aid contribute to social cohesion after civil war? Evidence from a field experiment in post-conflict Liberia. Am. Econ. Rev. 287–291.

Ferraz, C., Finan, F., 2007. Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes.

Field, E., Pande, R., Papp, J., Park, Y.J., 2012. Repayment flexibility can reduce financial stress: a randomized control trial with microfinance clients in India. Edited by Tiziana Leone PLoS One 7 (9), e45679. http://dx.doi.org/10.1371/journal.pone.0045679.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J.P., Allen, H., Baicker, K., Oregon Health Study Group, 2012. The Oregon health insurance experiment: evidence from the first year. Q. J. Econ. 127 (3), 1057–1106. http://dx.doi.org/10.1093/qje/qjs020.

Foote, C.L., Goetz, C.F., 2008. The impact of legalized abortion on crime: comment. Q. J. Econ. 123 (1), 407–423.

Franco, A., Malhotra, N., Simonovits, G., 2014. Publication bias in the social sciences: unlocking the file drawer. Science 345 (6203), 1502–1505. http://dx.doi.org/10.1126/science.1255484.

Fryer Jr., R., 2017. The production of human capital in developed countries: evidence from 196 randomized field experiments. In: Duflo, E., Banerjee, A. (Eds.), Handbook of Field Experiments, vol. 2, pp. 95–322.

Fujiwara, T., Wantchekon, L., 2013. Can informed public deliberation overcome Clientelism? Experimental evidence from Benin. Am. Econ. J. Appl. Econ. 5 (4), 241–255.

Giné, X., Karlan, D.S., 2014. Group versus individual liability: short and long term evidence from Philippine microcredit lending groups. J. Dev. Econ. 107, 65–83.

Glennerster, R., Powers, S., 2016. Balancing risk and benefit: ethical tradeoffs in running randomized evaluations. In: DeMartino, G.F., McCloskey, D.N. (Eds.), The Oxford Handbook of Professional Economic Ethics. Oxford University Press, Oxford, UK.

Glennerster, R., Takavarasha, K., 2013. Running Randomized Evaluations: a Practical Guide. Princeton University Press, Princeton, NJ.

Gueron, J.M., 2017. The politics and practice of social experiments: seeds of a revolution. In: Duflo, E., Banerjee, A. (Eds.), Handbook of Field Experiments, vol. 1, pp. 27–70.
Gueron, J.M., Rolston, H., 2013. Fighting for Reliable Evidence. Russell Sage Foundation, New York, pp. 1–22.
Haushofer, J., Shapiro, J., 2013. Household response to income changes: evidence from an unconditional cash transfer program in Kenya. Mass. Inst. Technol.
Herndon, T., Ash, M., Pollin, R., 2014. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Camb. J. Econ. 38 (2), 257–279.
Hicks, J.H., Kremer, M., Miguel, E., July 2015. Commentary: deworming externalities and schooling impacts in Kenya: a comment on Aiken et al. (2015) and Davey et al. (2015). Int. J. Epidemiol. http://dx.doi.org/10.1093/ije/dyv129 pii:dyv129.
Hoffmann, V., Barrett, C.B., Just, D.R., 2009. Do free goods stick to poor households? Experimental evidence on insecticide treated bednets. World Dev. 37 (3), 607–617.
Hoxby, C.M., 2000. "Does competition among public schools benefit students and taxpayers?". Am. Econ. Rev. 90 (5), 1209–1238. http://dx.doi.org/10.1257/aer.90.5.1209.
Hoxby, C.M., 2007. Does competition among public schools benefit students and taxpayers? reply. Am. Econ. Rev. 97 (5), 2038–2055. http://dx.doi.org/10.1257/aer.97.5.2038.
Humphreys, M., November 2, 2014. How to Make Field Experiments More Ethical. The Monkey Cage. https://www.washingtonpost.com/blogs/monkey-cage/wp/2014/11/02/how-to-make-field-experiments-more-ethical/.
Humphreys, M., Sanchez De La Sierra, R., Van Der Windt, P., 2012. Social and Economic Impacts of Tuungane: Final Report on the Effects of a Community Driven Reconstruction Program in Eastern Democratic Republic of Congo. Unpublished, Department of Political Science, Columbia University.
Hutton, J.L., 2001. Are distinctive ethical principles required for cluster randomized controlled trials? Stat. Med. 20 (3), 473–488. http://dx.doi.org/10.1002/1097-0258(20010215)20:3<473::AID-SIM805>3.0.CO;2-D.
Imbens, G., 2011. Experimental Design for Unit and Cluster Randomized Trials. International Initiative for Impact Evaluation (3ie), Washington, DC. http://cyrussamii.com/wp-content/uploads/2011/06/Imbens_June_8_paper.pdf.
Johnson, J., May 13, 2015. Campaign Experiment Found to Be in Violation of Montana Law. The Monkey Cage. https://www.washingtonpost.com/blogs/monkey-cage/wp/2015/05/13/campaign-experiment-found-to-be-in-violation-of-montana-law/.
J-PAL, 2015. Martin Hirsch/Government Panel: Creating Space for Evidence in Policymaking in France. https://www.youtube.com/watch?v=gCi60Dyxgws&list=PL5Dr5MK6NSso3iEqn6BDu8OzyMFyLwiNE&index=19.
Karlan, D., Appel, J., 2016. Failing in the Field: What We Can Learn When Field Experiments Go Wrong. Princeton University Press, Princeton, NJ.
Karlan, D., Zinman, J., 2010. Expanding credit access: using randomized supply decisions to estimate the impacts. Rev. Financ. Stud. 23 (1), 433–464. http://dx.doi.org/10.1093/rfs/hhp092.
Khan, A.Q., Khwaja, A.I., Olken, B.A., 2014. Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors, w206272014. National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w20627.pdf.
Klein, R.A., Ratliff, K.A., Vianello, M., Adams, R.B., Bahník, Š., Bernstein, M.J., Bocian, K., et al., 2014. Investigating variation in replicability: a 'many labs' replication project. Soc. Psychol. 45 (3), 142–152. http://dx.doi.org/10.1027/1864-9335/a000178.
Kling, J.R., Liebman, J.B., Katz, L.F., 2007. Experimental analysis of neighborhood effects. Econometrica 75 (1), 83–119.
Kremer, M., Glennerster, R., 2011. Improving health in developing countries. In: Handbook of Health Economics, vol. 2. Elsevier, pp. 201–315. http://linkinghub.elsevier.com/retrieve/pii/B9780444535924000049.
Kremer, M., Holla, A., 2008. Pricing and Access: Lessons from Randomized Evaluation in Education and Health. Citeseer.

Kremer, M., Miguel, E., 2007. The illusion of sustainability. Q. J. Econ. 122 (3), 1007–1065. http://dx.doi.org/10.1162/qjec.122.3.1007.
Leamer, E.E., 1983. Let's take the con out of econometrics. Am. Econ. Rev. 31–43.
McNutt, M., 2015. Editorial retraction. Science 348 (6239), 1100. http://dx.doi.org/10.1126/science.aac6638.
McRae, A.D., Weijer, C., Binik, A., Angela White, Grimshaw, J.M., Boruch, R., Brehaut, J.C., et al., 2011. Who is the research subject in cluster randomized trials in health research? Trials 12 (1), 183. http://dx.doi.org/10.1186/1745-6215-12-183.
Meager, R., August 2015. Understanding the Impact of Microcredit Expansions: a Bayesian Hierarchical Analysis of 7 Randomised Experiments. MIT Working Paper. http://economics.mit.edu/files/10595.
Miguel, E., Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. Econometrica 72 (1), 159–217. http://dx.doi.org/10.1111/j.1468-0262.2004.00481.x.
Miguel, T., 2015. "Introduction to Economics 270D." Presented at the Econ 270D: Research Transparency in the Social Sciences. University of California, Berkeley. http://emiguel.econ.berkeley.edu/assets/miguel_courses/12/Lectures-PDF.zip.
Mobarak, A.M., Rosenzweig, M., 2014. "Risk, insurance and wages in general equilibrium. Natl. Bur. Econ. Res.
Motl, J., 2014. Decision Finding Sufficient Facts to Demonstrate a Violation of Montana's Campaign Practice Laws. Commissioner of Political Practices of the State of Montana.
Muralidharan, K., 2017. Field experiments in education in the developing countries. In: Duflo, E., Banerjee, A. (Eds.), Handbook of Field Experiments, vol. 2, pp. 323–386.
Muralidharan, K., Sundararaman, V., 2011. Teacher performance pay: experimental evidence from India. J. Polit. Econ 111 (1), 39–77.
Olken, B.A., 2015. Promises and perils of pre-analysis plans. J. Econ. Perspect. 29 (3), 61–80.
Olken, B.A., Onishi, J., Wong, S., 2014. "Should aid reward Performance? Evidence from a field experiment on health and education in Indonesia. Am. Econ. J. Appl. Econ. 6 (4), 1–34. http://dx.doi.org/10.1257/app.6.4.1.
Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251), aac4716. http://dx.doi.org/10.1126/science.aac4716.
Ozler, B., October 15, 2014. How Scientific Are Scientific Replications? World Bank Blog. http://blogs.worldbank.org/impactevaluations/how-scientific-are-scientific-replications.
Ravallion, M., 2012. Fighting poverty one experiment at a time: a review of Abhijit Banerjee and Esther Duflo's *poor economics*: a Radical Rethinking of the Way to Fight global poverty. J. Econ. Lit. 50 (1), 103–114. http://dx.doi.org/10.1257/jel.50.1.103.
Research Network on Economic Equality & Social Interactions. MacArthur Foundation, n.d.
Ridker, P.M., Torres, J., 2006. Reported outcomes in major cardiovascular clinical trials funded by for-profit and not-for-profit organizations: 2000–2005. JAMA 295 (19), 2270. http://dx.doi.org/10.1001/jama.295.19.2270.
Rogoff, K., October 2013. FAQ on Herndon, Ash and Pollin's Critique of 'Growth in a Time of Debt'. Technical report. Unpublished Mimeo available on Rogoff's website at: http://tinyurl.com/ot8h53e.
Rogoff, K., Reinhart, C., 2010. "Growth in a time of debt. Am. Econ. Rev. 100 (2), 573–578.
Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86 (3), 638–641. http://dx.doi.org/10.1037//0033-2909.86.3.638.
Rothstein, J., 2004. Does competition among public schools benefit students and taxpayers? A comment on Hoxby (2000). SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.692582.
Rothstein, J., 2005. http://www.nber.org/papers/w11215.
Simonsohn, U., 2015. Small telescopes detectability and the evaluation of replication results. Psychol. Sci. 0956797614567341.
Thomas, D., E. Frankenberg, J. Friedman, J.-P. Habicht, M. Hakimi, N.J. Jaswadi, G. Pelto, B. Sikoki, T. Seeman, and J.P Smith. 2003. "Iron deficiency and the well-being of older adults: early results from a randomized nutrition intervention." In.

Vivalt, E., 2015. How Much Can We Generalize from Impact Evaluations? Unpublished Manuscript New York University.
Westfall, P.H., Young, S.S., 1993. Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.
William, J., Kremer, M., de Laat, J., Tavneet, S., 2016. Borrowing Requirements, Credit Access, and Adverse Selection: Evidence from Kenya (in press).
Zywicki, T.J., 2007. Institutional review boards as academic bureaucracies: an economic and experiential analysis. Northwestern Univ. Law Rev. 101, 861.