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

Economists have known for a long time that field experiments could help identify causal connections by solving the problem of selection bias, but the technique was not widely used outside a few specialized areas including welfare and labor market experiments in the US (an exception was Jamison et al. 1981). The dramatic increase in the use of the technique, first in development economics and now much more widely, was the result of a series of innovations which made the technique easier and cheaper to use. Some of these were theoretical. They included understanding how to maximize power from limited sample sizes (Imbens 2012; Bruhn and McKenzie 2009; see also Chapter xx) and how to use RCTs to measure externalities (Miguel and Kremer 2004), the diffusion of information (Duflo and Saez 2002; Kremer and Miguel 2007), equilibrium effects (Crepon et al. 2012), and parameters in network theory (Chandrasekhar, Kinnan, and Larreguy 2015; Beaman et al. 2013). But many of the innovations that powered the growth of field experiments 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 RCTs had been conducted in the past, these new partners tended to be more open to trying new approaches to solving problems and more willing to test different aspects of their programs separately and in combination. Their 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 A discusses how to select and work with a partner organization that will implement the program to be evaluated by an RCT as well as under what conditions it makes sense for a researcher to be both the implementer and the evaluator of a program. Section B examines a range of practical issues in measurement. It focuses in particular on how to create locally relevant, concrete indicators of abstract concepts such as empowerment and collective action. It also provides practical tips on keeping attrition rates low in panel surveys. Section C covers the practical ethical issues a researcher conducting randomized evaluations must take into account when designing and carrying out their research. Section D covers research transparency including avoiding publication bias through experimental registries and avoiding data mining through preanalysis plans.

A. Collaboration between researchers and implementers

Unlike most academic economic research, running field-based randomized control trials (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.

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?

What makes a good implementing partner?

i) Sufficient scale

A first and easy 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 XX) as well as the number of different variants of the program that are going to be compared, and the outcome of interest. Thus a lot of detailed discussion takes place about what a potential evaluation would look like before it is possible to say if an evaluation is feasible. However, 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.

ii) 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. As we discuss above, 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. The best partnerships are the ones in which researcher and implementer work together to decide what the most interesting versions of the program to test are.

iii) 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 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 are doing a proof-of-concept evaluation of the type discussed above. 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, noncash resources that would be hard to replace. 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, Duflo, Glennerster, and Kinnan 2015). They operated at a large scale and their credit product was close to that of many microcredit organizations. We tested their impact as they expanded into a large Indian city, a popular type of location for microcredit organizations.

iv) Local expertise and reputation

Implementers that 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 like, 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 l 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. As I discuss below, 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 in many papers 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.

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

v) Low staff turnover

There are many difficulties in working with governments and donor organizations, but perhaps the hardest to overcome is high staff turnover. As we have emphasized, evaluation is a partnership of trust and understanding and this takes time to build. All too often a key government or donor counterpart 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. The only way a researcher can protect the evaluation is to try and build relationships at many levels throughout the implementing organization, so that the loss of one champion does not doom the entire project. But this may not be sufficient. One of the many advantages of working with local NGOs is that they tend to have greater stability in their staffing.

vi) 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, page 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 till it is complete; transport costs of the program might be higher as implementation communities end up being further apart than they otherwise would be to allow for controls; roll-out plans could need to be set further in advance than normal to allow for the evaluation; selection criteria must be written down and followed scrupulously in order to reduce the discretion of local staff in accepting people into the program, and 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

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

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 project if left unchecked.

This commitment to the evaluation needs to be at many levels of the organization. If the headquarters in Delhi want to do an impact evaluation but the local staff don't, 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 implementer will need to participate in the study. Worst of all is when a funder forces an unwilling implementer to do an RCT run by a researcher. 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 does many types of programs has less at stake from an impact evaluation of one of their programs than an organization that has 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, Pande, Papp, and Park 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.

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 themselves a better partner?

i) Answer questions the partner wants 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 and include a component of the evaluation that answers the key questions of the implementer as well as elements that answer the key researcher questions. Sometimes these questions don't 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.

ii) Be flexible about the evaluation design

The research design a researcher has in their head when they start 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 twenty 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 that have some education but not too much. These concerns are relatively easy to deal with: agree to drop individuals or communities that don't 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 research 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. It 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 has been invested.

iii) 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 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 good will earned is invaluable when difficult issues arise later in the evaluation process.

iv) Provide intermediate outputs

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 endline 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 don't 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.

v) 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 to some time later to discover the results. Stuff 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.

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 test precisely the intervention or range of interventions we want to test. In order 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 that 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, Glennerster and Miguel 2012; Feron, Humphreys, and Weinstein 2009; Humphreys, Sanchez de la Sierra, and van der Windt 2012; and Beath, Christia, and Enikolopov 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 A vs that running B. 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 above, 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 efficient 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 a (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 (Field, Glennerster, Nazneen, Pimkina, and Sen, ongoing).[2]

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 don't 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, Berry, and Shapiro (2010) sought to understand whether price influenced use of health products (something that was not an issue for deworming pills) and 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

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

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.[3] Hoffman, Barrett, and Just (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 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 developed).

Researchers sometimes choose to work through research organizations as implementers or create new implementing organizations because their empirical and theoretical work suggests to them 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 Stick are examples where researchers helped set up new organizations to implement programs which were also evaluated through field experiments.[4]

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 well 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. I address the last point in the ethics section.

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 don't arrive on time, staff get sick or quit, 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 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.

[3] 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.

[4] 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.

To what extent can we generalize the results from researcher-led RCTs? Vivalt (2015), in a meta-analysis 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 Hoffman 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 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, 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 by 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 very 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 RCTs the benefits of self-implementation do not outweigh the costs. In particular, researchers isolate themselves from potentially very useful partners. We discuss later how to maximize the benefits from collaborating with implementing 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.[5]. I discuss this in the ethics section below.

B. Practical issues in measurement

Some of the biggest practical challenges in running randomized evaluation arise from the need to collect high quality data. There are some cases where researchers can rely on administrative data for their

[5] 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/.

outcomes and therefore do not have to collect their own (for example Angrist 1990; and Angrist, Bettinger, and Kremer 2006). In most cases, though, administrative data are unavailable to researchers, not collected on all individuals, are unreliable, or not detailed enough for the researcher's needs.

While the challenge of accurate measurement is not unique to field experiments, innovation in measurement has been a key driver of the expansion in the use of field trials and their expansion into new subject areas. Having a high-quality set of indicators can make or break a field experiment. The standard criteria for assessing the quality of an indicator are that they are logically valid, measureable, precise, and reliable. The definitions of these and how to assess indicators against them can be found in textbooks (e.g., Glennerster and Takavarasha, 2013). Here we focus on four specific practical challenges: developing indicators that measure not just whether but how an intervention has an impact; creating locally relevant, concrete indicators for abstract concepts such as collective action and empowerment; creating indicators that are not subject to social desirability bias; and practical ways to reduce attrition. We end by discussing ways to measure the quality of the data being collected and monitor the behavior of enumerators.

Indicators that explain the how--and not just whether--an intervention had an impact

A field experiment that provides evidence not just on whether an intervention led to a particular impact but also provides a description of the paths through which the impact was achieved is more interesting and more informative. This requires the researcher to develop a series of indicators that correspond to each section of a logical framework.

The following is a logical framework and associated indicators for a study that provided in-service training for teachers in order to improve their delivery of HIV-prevention education in primary schools in Kenya (Duflo, Dupas, and Kremer 2014) and comes from Glennerster and Takavarasha 2013. By having indicators at each step of the framework it is possible to tell, if the program is not effective, at which stage the logical chain broke: did the teachers get trained but not teach? Did the teachers teach but the children not learn the material? Or did the children learn the material but not change their behavior?

Table 1

	Objectives	Indicators
Input	Teachers are trained to offer HIV education	Hours of training implemented
Process	Teachers increase and improve HIV education	Hours of HIV education given; application of the program's teaching methods
Output	Students learn best practices for prevention	Test scores on HIV education exam

Impact	Students reduce their unprotected sexual behavior; incidence of HIV infection decreases	Self-reported sexual behavior; number of girls who have started childbearing; HIV status

Developing concrete indicators for abstract concepts such as trust and empowerment

Developing indicators for inputs and outputs is often easier than developing indicators for final outcomes. When the outcomes are relatively abstract concepts, developing appropriate indicators can be particularly challenging.

If we want to measure the impact a policy or program has on an abstract concept such as trust or empowerment we need to find an action or behavior which is the physical manifestation of that concept in the population in which the RCT is being conducted. The action or behavior needs to be one that is relevant for everyone in our population but that not everyone in our population engages in. In other words, there needs to be variation in the indicator in the population and the variation must be correlated with the underlying concept we are seeking to measure. If everyone engages in the behavior--or no one does--then it will not be feasible for a program or policy to achieve measurable change. These ideas are best explored through some examples.

Measuring collective action, and an example from Sierra Leone

A key challenge in measuring a concept like collective action is that it will be expressed differently in different communities. One community may come together to repair the leaking roof of a health clinic, another to pay for an additional teacher's aid in the local school, and a third to collectively market local agricultural produce. It is difficult, if not impossible, to reduce all of these different activities to a single indicator. We could try and ask about different collective activities people engage in and ask how much time people devote to each of them with the resulting single indicator being time spent on collective activity. One problem with this is being comprehensive. Without a lot of prompting, respondents may not realize that an activity they are undertaking is something the researcher would class as part of collective action. We would therefore need to go through a long list of possible activities with each respondent to find out if they take part in each. A similar problem occurs with questions about the number of groups a respondent is part of (a common measure of social capital): most respondents will not think of all the groups they are part of unless prompted with a pretty compete list of possibilities.

An alternative approach is to focus on a collective activity that all communities in the sample engage in to some extent. This allows us to measure not just whether the activity happens but the amount of the activity. It is much better to have a continuous measure than a dummy variable as this enables us to measure increases in the indicator across all the samples, not just those that start out at zero. If we have an activity that all communities engage in at some level then we can also collect data in all communities on who is involved in the activity. This allows us to look not only at the extent of collective action, but how broad the participation in it is.

Finding an activity that all communities participate in but which they participate in at different levels requires knowing the context well. Prior to the launch of the baseline survey for and evaluation of community driven development in Sierra Leone, Casey spent a year working with local partners to develop locally relevant indicators of collective action, trust, and participation (Casey, Glennerster, and Miguel 2012). One measure, which has since been used by other authors (e.g, Acemoglu, Reed, and Robinson 2014), was the extent of "road brushing." The combination of low population density, high rainfall, a lush tropical environment, and limited government funding for road maintenance means that the task of keeping dirt roads (or paths) between communities open and well maintained often falls to communities. The task is a clear public good as everyone benefits from being connected to other communities and the work is nontrivial, often including rebuilding bridges made of the trunks of palm trees. The indicator is continuous. Every community does some road brushing but some do it more frequently than others. It is also possible to measure how many in the community participate and how evenly the burden is spread. In a nationally representative sample 39 percent of rural households reported participating in road brushing in the last month with a standard deviation of 21 percent (NPS Surveys 2005 and 2007 reported in Glennerster, Miguel and Rothenberg, 2013). Clearly, frequency will be related to other characteristics such as remoteness but in an RCT these characteristics should be evenly spread between treatment and comparison so that a higher rate of roadbrushing in the treatment group would indicate and increase in collective action. The indicator has also been used to examine the relationship between collective action and ethnic diversity (Glennerster, Miguel, and Rothenburg 2013), and chiefdom competition and collective action (Acemoglu, Reed, and Robinson 2014).

Other common activities that communities come together to provide in Sierra Leone and other countries include support to community teachers and donations to the local school.

Measuring women's empowerment, and an example from Bangladesh

The second example comes from Field and Glennerster (ongoing), who wanted to measure the impact of a program on adolescent girls' empowerment in Bangladesh. The first step was to ask: What does an empowered girl in Bangladesh look like, and in what concrete ways is she different from an unempowered girl? This question was posed to a variety of stakeholders, including girls themselves and professionals working on empowerment programs. A number of themes were common to most of the answers. An empowered girl had more education, financial literacy and independent sources of income, mobility, knowledge about her health needs, and an ability to negotiate or advocate for her needs. Given the high rate of child marriage in Bangladesh, marrying after the age of 18 was seen as an important sign of empowerment.

The next step was to turn these subconcepts into concrete indicators. For some the task was reasonably straightforward. Greater education can be measured by years of schooling, highest class passed, and literacy and numeracy tests. For others it was more complicated. Even age of marriage, which might appear easy to measure, was complicated by a concern that some families would not admit to marriages

prior to age 18 given that such marriages are illegal in Bangladesh. Therefore data on girls' ages were collected prior to the start of the project and compared with date of marriage as a check on self-reported age of marriage.

To measure health knowledge the easiest approach is to ask specific questions that can be marked right or wrong. Even better is a question such as, "Name all the methods of contraception you know" because it produces a range of values that reflects a wider distribution of knowledge across the population. It is important to pitch these questions at the right level for the population. Ideally there will be some questions that many people get right, some that roughly 50 percent get right and some that only 10–20 percent get right. In general, a question that 5 percent or 95 percent of respondents respond to in the same way to is a waste of time on a survey. It is also important to make sure the questions reflect local health needs. In the area covered by this study, one-third of girls were stunted, so more of the health knowledge questions related to nutrition than they might in a study in a different population.

There is a tension in survey design between carefully tailoring the questions to the local context and also using indicators that are well known in the literature. There are two key benefits in using standardized indicators: one is that they have been thoroughly tested and validated; the other is that they allow our study sample to be compared to other samples. Thus, for the questions assessing mental health, Field and Glennerster used a series of standard questions even though some of these could have been better worded to reflect the local context, as the results could then be translated into an industry standard for depression. This made it possible to say whether the sample as a whole had higher or lower rates of mental health issues than other populations in other countries.

One of the hardest concepts to turn into a good indicator was mobility, although it was considered one of the key components of empowerment by those working in the area. Lack of mobility could restrict girls' ability to attend school, get the health care they needed, participate in community activities, engage in income-generating activities, and control their time and lives.

While mobility might appear to be mainly related to distance (e.g., how far a girl can travel) the answer turns out to be conditional on many other factors such as who accompanies her, what she is wearing, and what the objective of the travel is. This conditionality of the response to many interrelated factors is typical of some of the more abstract measures economists attempt to measure. For example, levels of trust in the community may be highly conditional on who is to be trusted with what. Someone may have high levels of collective action on issues of education but not on health (we discuss how to measure social capital above). It would be possible to ask a set of conditional questions that map out all these conditionalities. In this case: how far could a girl travel to see a relative if accompanied by her brother? How far could she travel to see a relative if accompanied by her father? But this would be very time-consuming and the results could be quite contradictory, with some girls scoring highly on some measures and poorly on others. Instead it is preferable to hone in on a couple of examples that are particularly pertinent to the population at hand. In our case, one of the motivators for looking at mobility was the concern that girls do not get the health care they need or the education they want because of restrictions on mobility. Here it turns out to be easier to simply measure the outcome: do girls manage to get an education and get treated when they are sick? Is it relevant that they are

accompanied to the school and clinic if they manage to receive the services? Is it a sign of lack of empowerment if adolescent girls who are sick are accompanied by their mothers to the clinic, or is this just a sign of a caring mother?

After several months of qualitative work it became apparent that while girls could often travel to school or to do tasks for others, many were restricted in their ability to travel to do something purely for their own enjoyment and that many girls chaffed against this restriction. This ability to do something for one's own enjoyment is arguably the essence of empowerment. Turning this into an indicator, however, required coming up with a specific activity that the vast majority of girls would want to do but only some had the ability to do. The final mobility section on the survey included two questions about going to the market outside the *para* (a cluster of houses, often with a common courtyard) to purchase a personal item. This was asked both with and without friends. It also included questions about going to meet friends to hang out. The final mobility section of the questionnaire is given below. The questions on dress, particularly wearing a more fashionable version of *salwa kamis*, again came from extensive qualitative interviews with girls who suggested they viewed this as a sign of an empowered girl and that wearing more fashionable *kamis* was something that many wanted to do but were restricted from doing.

Have you ever walked to school alone?
Are you allowed to go alone to a relative's house outside of the *para*?
Are you allowed to go alone to meet your friends outside of the *para* for any reason (to get school notes, hang out, etc.)?
Have you ever gone to the market outside of the *para* to buy personal items with friends (and without guardians)?
Have you ever gone to the market outside of the *para* to buy personal items alone?
Have you ever attended any sort of community events/activities outside of the *para*? (Ex: fair—Pehla Boishak, theater, cultural program, boat racing, religious event—Kitton, Puja, Mehfil)
Have you ever attended one of these events without guardians present (either alone or with friends)?
Do you always wear a burkha when leaving the house?
Would your parents give you permission to leave the house without wearing a burkha?
Each year there are new fashions that come out. If you wanted to wear some of these modern fashions (short *kurta/kamis*, *chos salawar*) and had the money to do so, do you think your parents would allow you?
Do you own a short *kamis* or *churida/chos paaa*-style *kamis*?

How many times in the past month have you played sports inside the *para*? (if haven't played, use code 00)	No. times ☐☐

Social desirability bias

People have a natural tendency to want to please and to be seen in a good light by those they are talking to. Either consciously or unconsciously this desire may influence how they respond to questions on a survey. This tendency is called social desirability bias. If both those in treatment and comparison consistently underrecord a behavior that is socially undesirable, it may still be possible to pick up the true impact of a program. If the underreporting is substantial, the true coefficient can be underestimated. Imagine, for example, that a program successfully reduced corruption but that no one reported any corruption: then it would be impossible to find an effect in the reported data.

The estimated coefficient can also be biased if the program being evaluated influences the degree to which people are reluctant to report the undesirable behavior. Adolescents who have gone through an HIV education program may be less willing to report multiple partners than the control group because they have just completed a program that stresses this is undesirable behavior. This bias may make the program appear effective even when it is not. In contrast, a program designed to encourage women to speak up about sexual abuse may lead to higher reporting in the treatment group even if the actual level of abuse suffered by those in the treatment group declined as a result of the program.

Ways to reduce social desirability bias include:

i) Asking about specific, time-bound activities

People may well provide an underestimate of the true level of an outcome when asked a general question such as, "How often do you miss school in any given month?" This systematic underreporting is less likely to happen when they are asked about a specific recent event: "Did you attend school yesterday?" Similarly, the answer to the question, "What is the main reason you miss school?" may be different from "Think back to the last time you missed school. What was the reason you missed school?"

ii) Discreetly observing behavior, rather than asking about behavior

When we are particularly concerned about social desirability bias we may not want to rely on self-reported behavior or attitudes. Very few people will admit in a survey to discriminating between people on the basis of their race or gender. However, it is all too common to observe discrimination in people's behavior. This may be because people are unaware of the subconscious biases they have or because they don't want to admit to them. The practical challenge in using observed behavior as an outcome for an RCT is that the behavior needs to happen at a time and place where we can observe it, and the situation needs to be as similar across respondents as possible. If, for example, we want to observe if respondents treat people of difference races differently, we may need to manufacture a situation in which they face people of different races. In order to rule out that the difference in behavior is due to other differences between people, such as age or gender or education, we need to manufacture a situation where they face people who are identical except for race. Bertrand and Mullainathan (2004) report on an experiment in which they sent manufactured curricula vitaes (CVs) in response to job applications. Different CVs were created and black- and white-sounding names were randomly attached to the CVs before they were submitted. Answering machines were set up to record whether a given

candidate was called back for interview. Otherwise identical CVs elicited 50 percent more callbacks if the candidate had a white-sounding name than if they had a black-sounding name.

Mystery riders, shoppers, or patients are another way to provide a structured observation of behavior. Keniston sent enumerators to bargain with rickshaw drivers in Jaipur and then record the different stages of the transaction (2011). Dizon-Ross, Dupas, and Robinson (2014) sent men into clinics in Ghana to see if they could purchase bed nets, which are meant to be given free to pregnant women and should not be sold. Das et al. (2008) had enumerators provide doctors with a set of symptoms and (afterwards) record what action the doctor recommended. In all these cases the respondent's job is to interact with people they do not necessarily know in a reasonably structured way. The researcher then adds even greater consistency to the interactions by having a fake client follow a script.

Observing a specific behavior that is not part of someone's job description is much harder. Observing group dynamics is harder still. However, it is possible to engineer even group situations and then observe the outcomes. Casey, Glennerster, and Miguel (2012) evaluate a community driven development program in Sierra Leone, a key objective of which was to change the dynamics of group decision making in communities. While community meetings are held frequently in these communities, they do not happen at specific times. Even if enumerators could find when one was planned, they would happen at different times and would discuss different issues, making it difficult to compare across meetings. The authors therefore created structured community activities. Specifically, when enumerators first went to communities as part of the endline they introduced themselves and asked for a meeting of the community to be organized. At the meeting the enumerators explained that in the following days they would be conducting interviews with a number of community members. To thank the community for their time the community was offered the choice between two gifts (packages of salt or batteries). The enumerators then recorded how the community came to the decision about which gift they preferred. By artificially prompting a real decision, the researchers were able to collect a wealth of data on behavior. How many people spoke at the meeting? How many women spoke? How many young people spoke? How long did the discussion last? Did a small group of elders break off to have a side conversation to settle the issue and then impose it on the wider community?

iii) Recording purchases

It may be relatively costless to claim to support a socially desirable outcome in a survey. How people spend their money may reflect underlying attitudes more accurately. If researchers want to track how money is spent, without relying on self-reports, they need to give respondents an incentive to purchase in a specific way. One approach is to offer a good for sale as part of the interview. For example, Ashraf, Berry, and Shapiro (2010) sell bottles of chlorine door-to-door at different prices and the outcome is purchases of the chlorine. Thornton (2008) sells condoms at the end of the survey. One problem with this is that people may feel obliged to purchase the good. This sense of obligation may or may not vary with treatment status. An alternative approach is to use vouchers that provide a discount on a purchase and have to be redeemed later. This separation in time and location may reduce the social pressure felt by the respondent to make a purchase to please the enumerator. Because the shopkeepers need to collect the subsidy they will return the vouchers to the researchers, making it possible to trace which

respondent used their voucher. Casey, Glennerster, and Miguel (2012) use vouchers as a way to measure collective action. Communities were given six vouchers that could be used at a local building supply store if they raised matching funds. Each voucher was worth the equivalent of US$17 if it was matched with US$33 of the communities' own money.

iv) Measuring a physical outcome

Another way to avoid relying on self-reported outcomes is to measure a physical outcome. For example, instead of asking about unprotected sexual intercourse we could test for the presence of sexually transmitted diseases. Instead of asking about whether people added chlorine to their water, we could test their water for chlorine. Instead of asking how much people exercised, we could weigh them. Such physical measures tend to be expensive and some are subject to noise. Weight, for example, is neither an exclusive nor exhaustive proxy for exercise as people can lose weight for other reasons, and people can exercise more and still not lose weight. However, finding a physical difference between treatment and comparison groups is a very convincing result because it addresses concerns of social desirability bias.

v) Using list experiments

In a list experiment, a respondent is asked how much of a list of activities they have engaged in, without being asked to identify whether they engaged in any specific activity. Usually most of the activities on the list are innocuous and unrelated to the field experiment, while the one the researcher is interested in is something people would be embarrassed to admit to. Because the treatment and comparison groups should engage in the innocuous activities equally, any difference between the average number of activities reported in the treatment and comparison group should be due to a difference in the activity the researcher is interested in. If some in the treatment and control group are given the list of activities without the inclusion of the item of interest, the researcher can know roughly how many in both treatment and comparison engage in the activity of interest. Karlan and Zinman (2010 and 2011) use this to assess how many people use microfinance loans for nonbusiness purposes even when the microcredit rules require the loan to be used for business reasons.

The approach allows the respondent to be truthful without the embarrassment involved in admitting to the behavior directly to an enumerator. This approach is only useful when the respondent is fully conscious of their behavior and thus is not useful if the respondent is unaware of their own bias. Also, this technique only provides estimates of the average number of times people in a group engaged in an activity. It cannot tell us who within the groups engaged in it. This means the approach cannot be used as an outcome in a regression that controls for individual level characteristics. One benefit of the approach is that it can be incorporated into a survey at relatively low cost, especially when compared to some of the other approaches discussed in this section.

vi) Implicit association tests

The human brain is quicker at sorting words and objects into categories if we have stronger associations between categories. Implicit association tests (IATs) use this to measure the extent of bias. Beaman and

coauthors (2009) use IATs to examine whether being exposed to a woman in a leadership role in India reduced bias against women leaders. Respondents were asked to sort pictures representing different objects or activities related either to leadership (e.g., a picture of the Indian parliament) or domestic activities (e.g., cooking) into two columns. In one configuration, a picture of a woman was placed on the same side as domestic activities needed to be sorted to, and a man on the side that leadership activities needed to be sorted to. In another configuration, the picture of the woman was placed on the side to which leadership activities should be sorted to and a picture of a man placed on the side to which domestic activities should be sorted to. Someone without gender bias would be as quick to sort in either configuration: the larger the difference in time to sort the two configurations, the greater the implicit gender bias.

A key advantage of IATs is that they can measure bias even when respondents are unaware that they are biased: a disadvantage is that they are expensive, time-consuming, and difficult to undertake.

vii) Using games

Recording how someone plays a game can reveal characteristics that a respondent might not want to or be able to articulate in a survey. By linking outcomes in the games to cash, researchers make it costly for respondents to deviate from their true preferences. There are a large number of different games that are designed to illuminate different characteristics of individuals. Perhaps the most common use of games in RCTs as a way to avoid social desirability bias is in the measurement of altruism. In the dictator game individuals are given some money and asked if they would like to pass on some of that money to another individual participating in the game. They are told that the other participant will never be told their identity. In some versions of the game the respondent is told some general facts about the person who will receive the money they give, such as their gender, race or economic status. This allows the researcher to see whether the respondent is more generous to those who are more similar to themselves.

Games are difficult and expensive to do well. They require researchers to set up strict protocols to ensure that confidentiality is maintained, that all subjects understand the choices they are making, and that the money that is promised goes to the right person. As the same games are often used in many different studies, it can be useful to rely on professional staff at a behavioral lab who are experienced at running games and are thus very familiar with the protocols. Permanent labs exist at many US universities and some European universities. There is even a permanent behavioral lab in Nairobi that researchers from around the world can apply to use (www.busaracenter.org). A downside of some permanent labs is that, being based in universities, most of their recruits are students who may be familiar with behavioral games and able to guess at researchers' objectives. As a result they may not respond to games in the same way as other segments of the population. If the objective is to use a game to measure an outcome of a field experiment it may be difficult to get the subjects to an appropriate lab.

Rao (2014) uses behavioral games to evaluate the impact of a policy that required elite private schools to accept some disadvantaged students. Rich students who had been exposed to more disadvantaged

students as a result of the policy shared more in dictator games when paired with poor students and when paired with other rich students.

For a fuller discussion of the range of behavioral games used by economists see Levitt and List (2007).

Practical tips for reducing 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:

i) Plan for more than one visit

Whether the surveys are conducted in people's homes or at 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 that the survey will take place on that day. Up to three separate visits may be needed to ensure that a high proportion of people are reached.

ii) 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 location. Baird, Hamory, and Miguel (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 through 2011 as they completed their education, married, and moved into the workforce. In the first round, 19 percent had moved out of the district and the team tracked respondents across Kenya as well as in Uganda, Tanzania, and even the UK.

iii) 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 US. 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.

iv) 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.

v) 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 (note that the respondents permission to get this data must be collected at baseline).

vi) 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.

vii) 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 will be terrible. 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.

viii) Consider an incentive

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 the survey, 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.

Measurement issues specific to RCTs

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 a RCT researcher has to be particularly concerned about: they all stem from the fact that some people in the sample will be exposed to an intervention (the treatment group) while others will not (the comparison group). It is essential to the validity of the experiment that data are collected in the same way in the treatment and comparison group and that the intervention does not interact 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 (Gine and Karlan 2014 use this approach when looking at microcredit contracts, and DeLaat et al. ongoing[6] use this to look at farming cooperative contracts--although both also 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.

The final measurement concern specific to impact evaluations is that the intervention being evaluated may have an effect on the relationship between the underlying outcome and the measurement of the outcome. This can occur if the intervention changes the incentive to misreport or has an impact on social desirability bias.

A particular problem is when a program provides an incentive to change a particular behavior. This incentive 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 Hana (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.

Measuring the quality of data collection

It is difficult to know how good or bad our data are unless we attempt to measure their quality. One way to do this is having a highly skilled enumerator (usually a supervisor) go back and re-ask some questions from a randomly chosen subsample of respondents and then compare the consistency of the two responses. Many researchers do "backchecks" of this kind on 10 percent of the sample. For larger surveys the rate can be lower than this, and for smaller surveys it may be worth doing backchecks for 15

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

percent of the sample. The backcheck survey does not have to be comprehensive. One reason for the backcheck is to make sure the enumerator is not making up data. Asking whether the respondent has been interviewed recently and asking simple questions where the respondent's answer is unlikely to change in the space of a few days are good for achieving this. Enumerators should be warned that backchecks will take place on an unannounced basis. It is good practice to make sure all enumerators have their work backchecked at least once in the first few days of a survey and any important discrepancies between the two surveys discussed with the enumerator. It should not be assumed 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.

Having enumerators take GPS readings from the interview site can help confirm that enumerators are going to the correct locations. Surprise visits from external monitors can also help assess quality and provide an incentive not to shirk. Electronic data collection now allows part of the interview to be recorded discreetly and at random as a check. It is also useful to look for systematic patterns in the data. There are usually important trigger questions in a survey which can change the length of the survey depending on the respondent's answer. In a demographic survey there will be many questions for each pregnancy a woman has had; in an agriculture 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 questions is a good way to spot poor-quality enumerators. These trigger questions are also important to check during the backcheck process.

C. Ethics[7]

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).

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.

[7] This section draws on Glennerster and Powers in *The Oxford Handbook of Professional Economic Ethics,* edited by George DeMartino (forthcoming).

46.101).[8] While the principles set out in the report were formulated in the US, they are reasonably general and are similar to the principles behind institutional review structures around the world.[9] 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.[10] 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

i) 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.

ii) Benefice

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.

iii) 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) bear the risks associated with research while another group receive the benefits. It recognizes that the individuals who take on the risks of research may not be precisely those who reap the benefits. Instead it aims to ensure that research is conducted amongst 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 benefice 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 to 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

[8] Accessed at http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.101, August 15, 2013.
[9] 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.
[10] PubMed, a database of medical research, reports over 325,000 medical trials registered between 1978 and 2013.

understands the risks and is willing to take them. But the right to be informed from the respect for persons is not absolute and is itself modified by the benefice 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 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.

Institutional review boards

As the principles make clear, there are difficult tradeoffs to make when determining the most ethical way to proceed with research. Researchers have the primary responsibility for judging the difficult trade-offs involved in conducting their research. However, they also have an interest in moving ahead with their research and this interest 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. Institutional review boards (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.[11]

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 by the Association for the Accreditation of Human Research Protection Programs (AAHRPP), 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 quite similar to the US. Australian research guidelines, for example, 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.

A surprising number of universities outside the US have no formal system of ethical review for research involving human subjects. In particular, because ethical review boards have mainly been seen as province of medical research, many universities that do not have medical schools do not have ethical

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

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 would be justifiable. Doctors and nurses have ethical obligations that go beyond research ethics—including to provide care to those in need in front of them. 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 they 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 (IFMR) 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, Das, and Rao 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.

One 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.

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 and has to seek approval both from their home university and from the Kenyan Medical Research Institute (KEMRI).

In the US, "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.[12]

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 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 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). 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 publically 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 published by the World Bank 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) in order to undertake their research, then approval is required even if they do not collect the data themselves.

[12] 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.

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

i) 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.

Complying with this requirement is usually relatively straightforward in the case of an experiment randomized at the individual level. 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 in the context of collecting baseline data and before randomization. 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.[13] 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 (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, Das, and Rao (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. 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

[13] 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.

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 below, 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 which participants cannot opt out of (Hutton 2001). Examples include adding chlorine to the community well, erecting street lights, or changing the rules under which the mayor is elected. 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 do not 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 type sof 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.

ii) 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 about the effectiveness of different forms of advertisements to reduce smoking amongst adults that randomized the position of antismoking billboards across the United States and then measured 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.

The other cost of collecting informed consent is that it could change a subjects' behavior to know they are part of an experiment. This change in behavior 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 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. In both cases a waiver is required from an IRB before the experiment can go forward. A researcher must justify the waiver 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 deception is warranted. IRBs will often require researchers to debrief subjects at the end of the experiment as a condition for gaining the waiver. One reason deceiving subjects about the real motivation of the research is frowned upon is that it creates negative externalities for other researchers: it may make subjects more suspicious of being involved in future research. This is more of an issue for lab experiments than field experiments because lab experiments are often run on students who may have participated in other experiments or know people who have. For more on deception and informed consent see Alderman, Das, and Rao (2013) and [add deception lab ref].

iii) 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 picture, is separated from the rest of the data as rapidly as possible; that only de-identified 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 consists of and how it was 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) 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 (like 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.

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 thus do research ethics govern the program or not? 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 in a footnote that they do not feel equipped to define the boundary between research and practice in social science. At the end of the section defining the separation of research and practice for medical research, footnote 3 states that:

> "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."

In 20xx a group was established to work on these guidelines but no additional guidelines were released. The practical question that faces researchers and IRBs evaluating research proposal from social scientists is if and when ethical approval should be sought for and research rules (including requirements for informed consent) applied to the program that is being evaluated. The discussion below 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 US 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 a 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 the 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 second cases the intervention was designed and implemented by someone else (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 it was the researcher who developed the drug who 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 an activity. 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) but notes that this "practice" falls 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 in order to create generalized knowledge. Thus if a program was to be rolled out by an NGO in a new area, this would not create generalized knowledge and would not count as research, and should be governed by the regulation of NGO activity rather than a research ethics board. However, if in order 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. 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 in order to generate knowledge. But as I have argued above, 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 have to get informed consent only from the individuals that 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.

The discussion surrounding the ethics of researcher implementation has become particularly heated recently because of a controversy surrounding a get-out-the-vote experiment conducted in Montana. While many of the facts of the case are not yet certain, it appears that the material used in the experiment included the state seal of Montana, which was not appropriate (as the material did not come from the State Government) and may even have been illegal. The appropriate content for election-related material is governed by law and researchers are subject to this law. But some have also suggested that it would be inappropriate if a research study influenced the outcome of an election. Presumably the objection applies only if researchers run the intervention, because researchers study interventions that influence elections all the time. But if interventions run by researchers should not change elections, does that mean that 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. One argument to suggest that elections are different is that while improving one person's health does not influence another person's health, election outcomes are a zero sum gain; a researcher cannot contribute to an overall improvement in society and instead must inevitably be helping one group at the expense of another. But many of the interventions that researchers study have distributional impacts as well as individual impacts. Is it unethical for a researcher to run a study that is designed to reduce school dropout of some children if those children then potentially cause a negative externality on other children whose job prospects are hurt by the improved education of their peers? 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. 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 standards or regulations from other implementers.

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

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 about is the treatment lottery (see Chapter x) in which some participants in the study are never given access to the program. The concern is that some potential participants in a program are "denied" access to the program in order to evaluate its impact. But 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 may aim to provide training to farmers in Tanzania but only has funding to cover two hundred farmers, far fewer than the number of all eligible farmers. A lottery is used to decide who receives access to the program but does not change the number of farmers 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 where a decision is taken to reduce the number of people who receive the program in the first phase in order to evaluate it. In this case the risk of harm is that the program is beneficial and delaying its introduction to all those who are 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 again 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. Ravalion (2012) suggests that allocating benefits randomly treats research subjects "merely as means to some end" and thus violates the respect-for-persons principle. But this concern only seems to be valid if we waive informed consent. True respect for persons means recognizing that people can make informed choices and may be willing to participate in an experiment that helps us learn what programs are and are not effective. There is nothing intrinsically unethical about allocation of a program being determined in part on the basis of evaluation needs as long as informed consent rules are followed.

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 1,000 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 1,000, 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 in 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 don't know if receiving the warm clothes is a benefit (if we did we would not be evaluating the program) and we usually don't 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 asset criteria for determining need are not necessarily the best way to identify need (Alatas, Banerjee, Hanna, Olken, and Tobias 2012). 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 amongst 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 above, instead of expanding the potential pool of families to 1,000 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.

If none of these options are workable and there is a high risk that the evaluation will lead to less good targeting of the program, this would not 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 treatment lottery around the 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 those who would have been just below the eligibility cutoff. Or the randomization can occur only amongst 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 into the program.

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. First, 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 that are 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).

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 risks 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.

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 amongst the qualified) but this has other risks: the evaluation cannot tell if the cutoff was too high, and it reduces access amongst 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 hence 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.

D. Transparency of research

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. Being transparent about the research process allows us to adjust the standard statistical tests to account for multiple hypotheses testing either within or between studies.

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 an overestimate (or underestimate) of 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 variance in our sample). 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 XX 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.

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).[14]

However, if we see only a select sample of the RCTs conducted, we may not draw a correct inference about the true effect size. In particular, 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). Figure [x] shows the case where the true effect size is zero and therefore the distribution of possible estimated effect sizes is centered around zero. Given *N*, the variance in the underlying population, there is a range of estimated effect sizes within which standard hypothesis testing would suggest had a greater than 5 percent probability of coming about by chance. There are three different RCTs and they provide estimated effect sizes $\hat{\beta}1$, $\hat{\beta}2$, and $\hat{\beta}3$. If we observe all three estimates we have a new estimated effect size based on all three studies which 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.

[add figure]

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 impact (in this case $\hat{\beta}2$) we may reject the null hypothesis, i.e., conclude that the true effect is different from zero with greater than 95 percent 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 [check] 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 above, only $\hat{\beta}3$ 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

[14] Economists rarely do a formal meta-analysis of this kind where coefficients are averaged in this way because we rarely see multiple RCTs of precisely the same intervention on the same population. Meta-analyses 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 meta-analysis as it is to a review of the literature.

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.

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.

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 i) record the existence of a study and the hypothesis it intends to test before the researcher (or journal) knows the results and ii) 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. While this two-step process is idea, the second step is harder to police, especially as some journals will not publish a study if the results have been reported elsewhere. 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 one 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 sort 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.

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 (RIDIE), 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 in order 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 register their trials.

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 at the start of the trial.

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 the publication bias section above. 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 creditably 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 an experimental research 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 from empirical economics 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 subject to whether or not outliers are dropped, and reviewers usually request authors to show that results are robust to including or excluding outliers. 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 vaccinations rates (valid measures include number of children with any vaccinations, proportion of children aged 2 to 5 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. As we discuss in section B, 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, Glennerster, and Miguel (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, in 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--while the true effect from a comprehensive examination of outcome indicators suggests a precise zero impact on any reasonable aggregate indicator of social capital. Each individual regression is valid. 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 call “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 which 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 which is the mean of all the individual asset dummies. We then estimate whether 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 Katz et al. (2001) in their evaluation of Moving to Opportunity is an alternative 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 a 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.

Using an index of mean effects can be used to reduce the number of outcome measures for which we estimate a treatment effect, but we may not want to aggregate all our outcome measures into a single family. We may, for example, want to examine the impact of the program on physical health, mental health, education, and employment separately. If we do this, we need to adjust the p-values for the fact that we are testing several related hypotheses. There are a number of approaches to adjusting p-values

for multiple hypothesis testing. The Bonferroni correction is the simplest of these, 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.

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 exactly the same issues as testing for multiple hypotheses, and we can adjust our p-values for multiple hypothesis testing in the same way.

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 pre-analysis plan (PAP). A PAP can be a useful complement to the strategies discussed above. 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 what 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 what subgroups we intend to test.

PAPs are becoming increasingly popular among those doing field experiments, but they are far from universal. Many authors feel they 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" like this can lead researchers to follow patterns that are there just by chance and that tying their hands in advance is useful. The most common use of PAPs is for field experiments where outcomes need to be constructed from many different individual indicators, 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. [expand on pros and cons of PAPs]

Writing a PAP is a time-consuming and difficult process as it requires the researcher to think through exactly how they plan to work through the analysis in the abstract, before the data are in front of them. If the researcher forgets to add something to the PAP, they can still do this analysis after they have the data, but it will have to be flagged as being outside the PAP and will lose credibility as a result. Thus it becomes critical to think through the entire analysis process carefully before finalizing the PAP.

The following elements should always be part of a PAP:

1. Which outcome measures will be primary and which secondary;

2. The precise definition of outcome variables including whether the composition of any indices or families that will be used for mean effects analysis;
3. Any subgroups that will be analyzed;
4. The direction of the expected impact if a one-sided test is to be used;
5. The precise specification to be used in the analysis, including whether controls will be included (and if so, which ones); whether the outcome variable will be in levels, logs or changes..

One issue that is still debated 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. For example, 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. 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 before the unblended data are examined (Olken 2013). [expand]

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. 2012), GoBifo (published as Casey et al. 2012) and the Oregon Health Insurance Experiment (published as Finkelstein et al. 2011). Since the opening of the AEA Registry, new PAPs in economics have been published at https://www.socialscienceregistry.org/.

Evidence on the magnitude of the problem

As discussed in the previous section greater transparency in research, in particular requiring pre-specification of analysis plans, has costs as well as benefits. In this section we discuss evidence on the magnitude of the biases that may result from lack of transparency.

Estimating the impact of lack of transparency is a challenge because we cannot observe the results of studies that do not get released or the results of alternative specifications that researchers run but which they do not report in their published study. We have to infer from what we see, whether this is in fact the whole universe of studies and estimations.

Other disciplines have examined the extent of bias in different ways. In the medical literature, where there are often many studies on a given intervention or drug, a correlation has been found between who funds an experiment and the magnitude of the effect found in the experiment suggesting that researcher manipulation has real impacts. In the psychology literature, where studies can often be run relatively quickly and cheaply, attempts to estimate the magnitude of bias due to lack of transparency

have focused on efforts to replicate results of studies in other labs. The lack of replicability of some important psychology results has raised concern that the level of bias is considerable.

Replication efforts can also 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 replication 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 (2014) 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 a the new study fails to replicate the first study. He suggests that the appropriate standard is whether the replication results are consistent 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.

Brodeur et al. (2013) 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 percent, or if studies with results under these levels are less likely to be published there will be few published results just below these cut offs. By examining empirical studies form three top journals (AER, JEP and QJE) between 2005 and 2011 Brodeaur 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. Brodeaur et al. estimate that 10–20 percent of p-values below 0.05 should in fact be between 0.10 and 0.25. Olken points out that this means that of 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.

References

Acemoglu, Daron, Tristan Reed, and James A. Robinson. 2014. “Chiefs: Economic Development And Elite Control Of Civil Society In Sierra Leone.” *Journal Of Political Economy* 122 (2): 319-368. doi:10.1086/674988.

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias. 2012. “Targeting The Poor: Evidence From A Field Experiment In Indonesia.” *American Economic Review* 102 (4): 1206-1240. doi:10.1257/aer.102.4.1206.

Alderman, Harold and Das, Jishnu and Rao, Vijayendra, “Conducting Ethical Economic Research: Complications from the Field.” World Bank Policy Research Working Paper No. 6446, 2013. Available at SSRN: http://ssrn.com/abstract=2265757

Anderson, Michael L. 2008. “Multiple Inference And Gender Differences In The Effects Of Early Intervention: A Reevaluation Of The Abecedarian, Perry Preschool, And Early Training Projects.” *Journal Of The American Statistical Association* 103 (484): 1481-1495. doi:10.1198/016214508000000841.

Angrist, Joshua. 1990. “Lifetime Earnings And The Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records.” *American Economic Review* 80 (3): 313-336.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. “Long-Term Educational Consequences Of Secondary School Vouchers: Evidence From Administrative Records In Colombia.” *American Economic Review* 96 (3): 847-862. doi:10.1257/aer.96.3.847.

Ashraf, Nava, James Berry, and Jesse M Shapiro. 2010. “Can Higher Prices Stimulate Product Use? Evidence From A Field Experiment In Zambia.” *American Economic Review* 100 (5): 2383-2413. doi:10.1257/aer.100.5.2383.

Ashraf, Nava, James Berry, and Jesse M Shapiro. 2010. “Can Higher Prices Stimulate Product Use? Evidence From A Field Experiment In Zambia.” *American Economic Review* 100 (5): 2383-2413. doi:10.1257/aer.100.5.2383.

Baird, Sarah, Joan Hamory, and Edward Miguel, "Tracking, Attrition and Data Quality in the Kenyan Life Panel Survey Round 1 (KLPS-1)." Center for International and Development Economics Research Working Paper Series, Center for International and Development Economics Research, Institute for Business and Economic Research, UC Berkeley, 2008.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. “The Miracle Of Microfinance? Evidence From A Randomized Evaluation.” *American Economic Journal: Applied Economics* 7 (1): 22-53. doi:10.1257/app.20130533.

Barrett, Christopher, and Michael Carter. 2014. “Retreat From Radical Skepticism: Rebalancing Theory, Observational Data And Randomization In Development Economics” In *Field Experiments And Their Critics*, 1st ed., 58-78. New Haven: Yale University Press.

Beaman, Lori, Niall Keleher, and Jeremy Magruder. 2013. “Do Job Networks Disadvantage Women? Evidence From A Recruitment Experiment In Malawi.” Working paper, Northwestern University, Evanston, Illinois, 2013. http://are.berkeley.edu/~jmagruder/recruitment.pdf.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. “Powerful Women: Does Exposure Reduce Bias?” *Quarterly Journal Of Economics* 124 (4): 1497-1540. doi:10.1162/qjec.2009.124.4.1497.

Beath, Andrew, Fotini Christia, and Ruben Enikolopov. 2013. “Democratization, Division Of Responsibilities And Governance Quality: Experimental Evidence On Local Institutions In Afghanistan.” *SSRN Journal*. doi:10.2139/ssrn.2202563.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily And Greg More Employable Than Lakisha And Jamal? A Field Experiment On Labor Market Discrimination.” *American Economic Review* 94 (4): 991-1013. doi:10.1257/0002828042002561.

Bruhn, Miriam, and David McKenzie. 2009. “In Pursuit Of Balance: Randomization In Practice In Development Field Experiments.” *American Economic Journal: Applied Economics* 1 (4): 200-232. doi:10.1257/app.1.4.200.

Casey, K., R. Glennerster, and E. Miguel. 2012. “Reshaping Institutions: Evidence On Aid Impacts Using A Preanalysis Plan.” *Quarterly Journal Of Economics* 127 (4): 1755-1812. doi:10.1093/qje/qje027.

Chandrasekhar, Arun, Cynthia Kinnan, and Horacio Larreguy, “Social Networks As Contract Enforcement: Evidence From A Lab Experiment In The Field.” NBER Working Paper No. 20259, 2015.

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. “Women As Policy Makers: Evidence From A Randomized Policy Experiment In India.” *Econometrica* 72 (5): 1409-1443. doi:10.1111/j.1468-0262.2004.00539.x.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, “Do Labor Market Policies Have Displacement Effects? Evidence From A Clustered Randomized Experiment.” NBER Working Paper No. 18597, 2012.

Das, Jishnu, Jeffrey Hammer, and Kenneth Leonard. 2008. “The Quality Of Medical Advice In Low-Income Countries.” *Journal Of Economic Perspectives* 22 (2): 93-114. doi:10.1257/jep.22.2.93.

Dhaliwal, Iqbal, and Rema Hanna, “Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India.” National Bureau of Economic Research Working Paper No. w20482, 2014.

Dizon-Ross, Rebecca, Pascaline Dupas, and Jonathan Robinson. “Governance And The Effectiveness Of Public Health Subsidies.” Working paper, Stanford University, Palo Alto, CA, 2014. http://web.stanford.edu/~pdupas/Governance&Effectiveness_PublicHealthSubsidies.pdf.

Duflo, Esther, and Emmanuel Saez. 2002. “Participation And Investment Decisions In A Retirement Plan: The Influence Of Colleagues’ Choices.” *Journal Of Public Economics* 85 (1): 121-148. doi:10.1016/s0047-2727(01)00098-6.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, “Education, HIV, And Early Fertility: Experimental Evidence From Kenya.” NBER Working Paper No. 20784, 2014.

Fearon, James, Macartan Humphreys, and Jeremy M. Weinstein, “Development Assistance, Institution Building, And Social Cohesion After Civil War: Evidence From A Field Experiment In Liberia.” Center For Global Development Working Paper No. 194, 2009. http://www.columbia.edu/~mh2245/papers1/CGD_WP.pdf.

Field, Erica, Rachel Glennerster, Shahana Nazneen, Svetlana Pimkana, and Iman Sen. 2015. “The Effect Of Financial Incentives And Girls Empowerment Curriculum On Adolescent Marriage And Schooling In Rural Bangladesh: A Clustered Randomized Controlled Trial.” Working paper, Duke University, Durham, North Carolina, 2015.

Field, Erica, Rohini Pande, John Papp, and Y. Jeanette Park. 2012. “Repayment Flexibility Can Reduce Financial Stress: A Randomized Control Trial With Microfinance Clients In India.” *Plos ONE* 7 (9): e45679. doi:10.1371/journal.pone.0045679.

Giné, Xavier and Karlan, Dean S., Group Versus Individual Liability: Long Term Evidence from Philippine Microcredit Lending Groups. Yale Economics Department Working Paper No. 61; Yale University Economic Growth Center Discussion Paper No. 970, New Haven, 2009. Available at SSRN: http://ssrn.com/abstract=1407614

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton, New Jersey: Princeton University Press.

Glennerster, Rachel, Edward Miguel, and Alexander D. Rothenberg. 2013. “Collective Action In Diverse Sierra Leone Communities.” *The Economic Journal* 123 (568): 285-316. doi:10.1111/ecoj.12030.

Haushofer, Johannes, and Jeremy Shapiro. “Household Response To Income Changes: Evidence From An Unconditional Cash Transfer Program In Kenya.” Working paper, Princeton University, Princeton, NJ, 2013. http://www.princeton.edu/~joha/publications/Haushofer_Shapiro_UCT_2013.pdf

Hoffmann, Vivian, Christopher B. Barrett, and David R. Just. 2009. “Do Free Goods Stick To Poor Households? Experimental Evidence On Insecticide Treated Bednets.” *World Development* 37 (3): 607-617. doi:10.1016/j.worlddev.2008.08.003.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2015. “Social And Economic Impacts Of Tuungane Final Report On The Effects Of A Community Driven Reconstruction Program In Eastern Democratic Republic Of Congo.” Working paper, Columbia University, New York, 2015. http://www.oecd.org/countries/democraticrepublicofthecongo/drc.pdf.

Hutton, J. L. 2001. “Are distinctive ethical principles required for cluster randomized trials?” *Statistics in Medicine* 20(3): 473-88.

Imbens, Guido, and Michal Kolesar. “Robust Standard Errors In Small Samples: Some Practical Advice.” NBER Working Paper 18478, 2012.

Jamison, Dean T., Barbara Searle, Klaus Galda, and Stephen P. Heyneman. 1981. “Improving Elementary Mathematics Education In Nicaragua: An Experimental Study Of The Impact Of Textbooks And Radio On Achievement.” *Journal Of Educational Psychology* 73 (4): 556-567. doi:10.1037/0022-0663.73.4.556.

Karlan, Dean, and Jonathan Zinman. 2010. “Expanding Credit Access: Using Randomized Supply Decisions To Estimate The Impacts.” *Review Of Financial Studies* 23 (1): 433-464. doi:10.1093/rfs/hhp092.

Karlan, Dean, and Jonathan Zinman. 2011. “Microcredit In Theory And Practice: Using Randomized Credit Scoring For Impact Evaluation.” *Science* 332 (6035): 1278-1284. doi:10.1126/science.1200138.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. "Moving To Opportunity In Boston: Early Results Of A Randomized Mobility Experiment." *Quarterly Journal of Economics* 116(2): 607-54.

Keniston, Daniel. “Bargaining And Welfare: A Dynamic Structural Analysis.” Working paper, Massachusetts Institute of Technology, Cambridge, MA, 2011. http://isites.harvard.edu/fs/docs/icb.topic983517.files/Keniston%20Bargaining_Write-up_Sept2011_11.11.01.pdf.

Kremer, Michael and Edward Miguel. 2007. “The Illusion Of Sustainability.” *Quarterly Journal Of Economics* 122 (3): 1007-1065. doi:10.1162/qjec.122.3.1007.

Leamer, Edward. 1983. “Let's Take the 'Con' Out of Econometrics.” *American Economic Review* 73*(1):* 31-43.

Levitt, Steven D, and John A List. 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal About The Real World?” *Journal Of Economic Perspectives* 21 (2): 153-174. doi:10.1257/jep.21.2.153.

McRae, Andrew D., et al. 2011. “Who is the research subject in cluster randomized trials in health research?” *Trials* 12(183). doi:10.1186/1745-6215-12-183

Miguel, Edward, and Michael Kremer. 2004. “Worms: Identifying Impacts On Education And Health In The Presence Of Treatment Externalities.” *Econometrica* 72 (1): 159-217. doi:10.1111/j.1468-0262.2004.00481.x.

Rao, Gautam, “Familiarity Does Not Breed Contempt: Diversity, Discrimination And Generosity In Delhi Schools.” Working paper, Harvard University, Cambridge, MA, 2014. http://scholar.harvard.edu/rao/publications/familiarity-does-not-breed-contempt-diversity-discrimination-and-generosity-delhi.

Ravallion, Martin. 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.” *Journal Of Economic Literature* 50 (1): 103-114. doi:10.1257/jel.50.1.103.

Rosenthal, Robert. 1979. “The File Drawer Problem And Tolerance For Null Results.” *Psychological Bulletin* 86 (3): 638-641. doi:10.1037//0033-2909.86.3.638.

Thomas, Duncan, and Elizabeth Frankenberg. 2002. “Health, Nutrition And Economic Prosperity: A Micro-Economic Perspective.” *Bulletin Of The World Health Organization* 80: 106-113.

Thornton, Rebecca L. 2008. “The Demand For, And Impact Of, Learning HIV Status.” *American Economic Review* 98 (5): 1829-1863. doi:10.1257/aer.98.5.1829.

Vivalt, Eva. “How Much Can We Generalize from Impact Evaluations?” Working paper, New York University, New York, 2015. http://evavivalt.com/wp-content/uploads/2014/12/Vivalt-JMP-latest.pdf

Westfall, P.H. and S.S. Young. 1993. *Resampling-based multiple testing: examples and methods for P-value adjustment.* New York: John Wiley and Sons.

Zywicki, Todd J. 2007. “Institutional Review Boards as Academic Bureaucracies: An Economic and Experiential Analysis.” George Mason Law & Economics Research Paper No. 07-20. Available at SSRN: http://ssrn.com/abstract=983649