Balancing Risk and Benefit: Ethical Tradeoffs in Running Randomized Evaluations

Rachel Glennerster and Shawn Powers

1. Introduction

There has been a dramatic rise in the use of randomized evaluations, particularly in development economics, in the last twenty years. This research has tested everything from the provision of additional textbooks in schools to providing information to voters on their elected officials. With the rise of the use of this methodology there have come questions about how and when the approach can be used ethically.

While the use of random assignment to answer social and economic questions is increasingly common, it is not new. An early example examined questions of energy policy in 1966 (quoted in Levitt and List 2009). This was followed by a series of negative income tax experiments in the US starting in the late 1960s and the Rand Health experiment completed in 1982. There have, however, been changes in how the tool has been used: in particular, researchers have increasingly engaged in intense collaborations with small-scale implementers, who can be more flexible and responsive than traditional government partners. There have been practical and methodological innovations which have introduced new ways to randomize, enabled us to measure spillovers, improved our ability to measure outcomes such as corruption and empowerment, and helped us maximize statistical power under tight budget constraints (see for example Miguel and Kremer 2004; Olken 2007; Bruhn and McKenzie 2009; Abadie and Imbens forthcoming; and Barios et al. 2010). One consequence of the innovations in measurement is that random assignment has been used to examine questions in new subject areas, such as women's empowerment (Chattopadhyay and Duflo 2004) and building trust in post-conflict communities (Casey, Glennerster, and Miguel 2012).

This work has not taken place in an ethics regulation vacuum. In 1974, the United States put into place a framework for research on medical and non-medical research involving human subjects. The ethical guidelines produced under this framework (the Belmont Principles) are similar to other ethical regulatory guidelines around the world, such as the Tri-Council Policy Statement in Canada and the National Statement on Ethical Conduct in Human Research in Australia.

We begin this chapter with a discussion of the Belmont Principles because we consider them a sound basis for judging the ethical issues we will discuss later (Section 2). Next we discuss whether and in what dimensions randomized evaluations of social programs raise different ethical questions from other research and evaluations and other forms of program provision (Section 3).

In Section 4 we discuss some of the practical issues that arise in applying the Belmont Principles to running randomized evaluations, especially in developing countries. We include in this discussion issues that are common to randomized and nonrandomized research. Throughout this chapter we draw on the experience and examples of researchers affiliated with our organization, the Abdul Latif Jameel Poverty Action Lab (J-PAL), who have collectively implemented hundreds of randomized evaluations in developing and developed countries. However, the opinions expressed here are our own and do not necessarily represent a consensus of all J-PAL affiliated researchers or the Massachusetts Institute of Technology.

A theme throughout the chapter is the importance of balancing risks and benefits: a trade-off that is central to the ethical guidelines enshrined in the Belmont Report. Because researchers may have an incentive to see the trade-off in a way favorable to the continuation of their research, it is important to have well-functioning institutions to oversee the assessment of risks and benefits. In the final section we discuss the extent to which these systems are in place for social sciences internationally and the ways in which they could be improved.

2. A Framework for Thinking About the Ethics of Randomized Evaluations

The Belmont Report, which was issued in 1978 by the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, provides the basis for decisions about the ethics of research carried out by research funded by most federal departments or agencies (Code of Federal Regulations, title 45, sec. 46.101).[1] Most US universities draw on this report and subsequent guidance from the Office for Human Research Protections (OHRP) to establish human subject procedures covering all research carried out by their faculty and staff, whatever the location or funding source.[2]

While the principles set out in the report were formulated in the US, they are reasonably general and are the basis of many other institutional review structures around the world.[3] 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.[4] While medical research, with the attendant level of potential risk, was clearly in the mind of those writing the principles, the principles explicitly cover nonmedical studies and recognize that the level of scrutiny and safeguards should be adjusted to the level of risk.

[1] Accessed at http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.101, August 15, 2013.
[2] Entities that do research outside of universities have adopted a variety of approaches to ethical oversight. Some organizations, such as Innovations for Poverty Action and Abt Associates, maintain their own internal IRBs, which follow OHRP standards. 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.
[3] 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.
[4] PubMed, a database of medical research, reports over 325,000 medical trials registered between 1978 and 2013.

We start by briefly describing the main principles, the ethical traditions on which they are based, and how they related to principles in economics, in order to discuss how and when randomized evaluations are compatible with these guidelines.[5] The three principles set out in the Belmont Report are:

- Respect for Persons: people's right to make their own decisions must be respected. In particular, participants in research must give informed consent to participate in the research. As discussed in Section 4, in some cases where the risks are minimal and the costs high, the requirement for informed consent can be waived. If the risks of a particular study are high, however, more detailed procedures may be required to ensure that participants fully understand the risks before giving informed consent. Those who find it hard to understand risks or cannot easily refuse to participate (like prisoners) need extra protection. Providing very high levels of incentives to participate in a study could potentially undermine a subject's ability to assess risk, requiring careful oversight.
- Beneficence: researchers should seek to increase people's well-being and avoid knowingly doing harm, for example by subjecting participants to unnecessary and painful tests or increased stress. However, avoiding all risk can be harmful to society in general if it prevents research that could have widespread benefits. Thus the overall benefits of the research to society always need to be balanced against the risks. Researchers should seek to maximize possible benefits and minimize possible harm.
- Justice: there should be fairness in the allocation of risks and benefits between different groups of people. It is important to avoid a situation where one group bears all the risk and another stands to reap all the benefits. For example, it would be unjust if a potentially risky vaccine were tested on prisoners and once approved was only made available to the rich.

The principles reflect a melding of two different traditions in ethics: a rights-based approach and a utilitarian approach. The respect for persons principle reflects the view that people should have the right to pursue their own happiness and make their own choices, including about risks and benefits. Welfare economics, which comes more from the utilitarian tradition, similarly enshrines a respect for individual preferences and finds that (under conditions such as full information and no externalities), allowing individuals to make their own choices maximizes individual utility. The practical implication is that it is better to regulate the provision of information to study participants than to regulate whether someone can participate in a program or study.

Nevertheless, the Belmont Principles do allow for exceptions to this freedom to participate because of the fear that some individuals may not assess risk rationally. This perspective has echoes in behavioral economics, which finds that people tend to weigh risks inconsistently, "overweighting" short-term incentives and "underweighting" long-term risks.

The beneficence principle comes squarely from a utilitarian tradition in explicitly discussing trade-offs between risks and benefits and thus has strong parallels with welfare economics. The principle recognizes that there may be cases where research benefits society as a whole even though risks may be taken by a few individuals, much as an economist would discuss the concept

[5] The Commission was established in the US following concerns about unethical research conducted during and after the Second World War. The report can be accessed from the website of the US Department of Health and Human Services: http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

of maximizing expected social welfare. While regulatory systems have increasingly relied on more rigid rules for assessing the ethics of research in a way which is much more in line with a rights-based approach than a utilitarian one, the Belmont Report and Common Rule nearly always include a caveat that where the benefits outweigh the costs, waivers can be given.

Simple utilitarianism is sometimes criticized for not taking the equitable distribution of costs and benefits into account. As a result, many ethical systems put some constraints on a simple criterion of maximizing aggregate social welfare. In the Belmont principles this comes in two forms. First, as discussed above, respect for persons means that the individual bearing the risk can decide to opt out of research, even though it may be beneficial to society as a whole. Second, the justice principle addresses the issue of equitable distribution directly. The justice principle has a parallel in the economic concept of Pareto improvement. While a policy may maximize social welfare, it is not a Pareto improvement if it increases total welfare by raising benefits for some while imposing costs on others. In practice, there are virtually no situations where we can guarantee that no one will be made worse off by an action or policy change. Thus Pareto improvement is rarely a criterion for judging policy options. Nor does the justice principle go as far as requiring Pareto improvement; it does not say that each individual must, in expectation, gain. Instead, the principle looks at groups: for example, if women in Ethiopia take risks for the study, then women in Ethiopia should (in expectation) gain from the results of the study.

Applying these principles is not always straightforward, and in practice research poses trade-offs among different ethical claims. For example, the Belmont Report notes that finding effective treatments for childhood diseases justifies medical research involving children, even when the children in a particular study do not benefit directly. However, when research presents "more than a minimal risk without immediate prospect of direct benefit to the children involved," then different claims under the principle of beneficence come into conflict. These trade-offs require value judgments. For this reason researchers in many institutions must answer to Institutional Review Boards (IRBs), so that the researchers themselves are not the last word on the ethics of a study. In the discussion that follows, we attempt to clarify the ethical trade-offs involved in RCTs, but the fact that there are so many trade-offs underscores the importance of having an appropriate institutional framework that can balance risks and benefits in an objective way. We return to questions about how well the existing institutional frameworks (particularly outside the US) operate at the end of this chapter.

3. What is Different, Ethically, About Randomized Evaluations?

As we noted in the introduction, the increasing use of randomized evaluations in economics has been associated with an increase in discussion about ethical issues. As we discuss specific ethical questions that arise during the conduct of some randomized evaluations, it is worth understanding what is different, and what is not different, ethically, about RCTs.

To answer this question we need to consider the counterfactual: what are we judging RCTs against? In what follows, we have to consider both what is different about the evaluation but also what is different about the implementation of the program because of the involvement of researchers. We focus on four aspects of randomized evaluations that have ethical implications: their potential to answer policy-relevant questions by cleanly identifying causal impact;

collection of primary data on individuals and communities; close collaboration between researchers and implementers; and the randomized methodology itself, which may change how programs are delivered.

a. Potential to answer policy-relevant questions by cleanly identifying causal impact

The ability to cleanly identify causal impact is a key reason that randomized evaluations have become an increasingly popular tool of economists and social scientists. There is surprisingly little high-quality evidence on the effectiveness of alternative ways to reduce poverty. Often the only evidence we have is based on comparisons of outcomes before and after the program started. However, this type of time series evidence implicitly assumes that all the changes that happen over time are the result of the program, which is a strong assumption. Another common form of evidence used by practitioners to inform program design and policy is comparisons of outcome between program areas and non-program areas. But programs are often introduced in particular areas at particular times for specific reasons. For example, a new education initiative may be piloted in schools with enthusiastic or motivated principals. Outcomes in these schools may exceed those in other schools for reasons that have more to do with the motivated principal than the particular program. In contrast, if a program is randomly assigned to a particular set of schools, communities, or individuals, we can be confident that the only systematic difference between the treatment and comparison group is that one has access to the program and the other does not. With sufficient sample size, this allows us to measure the causal impact of the program on outcomes.

Randomized evaluations are not the only methodology that allows social scientists to identify the causal impact of a policy and program. Quasi-experimental approaches such as instrumental variables or regression discontinuity can be very effective methodologies in the right situation. However, the use of quasi-experimental approaches is often constrained by the need to find a quirk that leads to near-random variation in how a program is implemented. For example, regression discontinuity designs can only be used if a program has a cutoff score for deciding eligibility and this is strictly enforced. [6] Randomized evaluations, in contrast, can be designed explicitly to answer very specific questions of importance to those designing antipoverty programs.

Why is this relevant to ethical considerations? Because risks and benefits of research always have to be balanced. The more likely a research project is to be able to answer a question in an unbiased way, and the more important that question is to designing more effective policy, the higher the benefit to conducting the research for a given risk, or similarly the more risk it is

[6] A regression discontinuity design is an empirical technique that exploits a discontinuity based on a cutoff value of a continuous variable. For example, a means-tested program may provide support to people below a given threshold of income, and no support to those above the threshold. This creates an opportunity to estimate the program's impact: those immediately above and below the cutoff are likely to be quite similar (provided that the variable used in the means test is not easy for people to manipulate), except that one group receives the program and the other does not. In this example, involving a researcher may cause the means test to be enforced more rigorously than it would have been in the absence of an evaluation.

acceptable to take. Of course, just because a study generates lots of benefit does not mean we should take on more risk for the sake of it. The Belmont Principles are clear that researchers should seek to minimize risk where they can. But it is ethical to take on some risk if it is necessary to achieve the benefit, and good identification of the causal impacts of relevant programs is an important benefit.

Critics of randomized evaluations will argue, however, that while randomized evaluations answer questions in an unbiased way, they may be too small scale and too specific in the questions they ask to provide useful information to those wanting to design more effective policy. Again, this is relevant to ethics: if the information generated is not of wide benefit to society then it is harder to justify any risk of harm involved in the evaluation. This contention, articulated for example by Ravallion (2008) and Deaton (2010), rests on the suggestion that understanding the impact of one program in one context at one time may be uninteresting because there is little reason to believe the results will generalize to other contexts. It is fine to test the effectiveness of a vaccine on a specific population and assume that it will generalize to others, but the way in which people respond to social and economic programs is so mediated by local institutions that generalization from one small context to another is hard.

One response is to run RCTs on large representative samples. For example, a randomized evaluation of teacher performance pay in Andhra Pradesh, India, tested the program in a representative sample of school in a random sample of districts across the state, which has a population of 84 million people (Muralidharan and Sundararaman 2011). It can be expensive to run an RCT on a randomly selected representative sample of a larger population, as it involves a much more dispersed study population. As always, these costs need to be compared to the benefits. But it is good practice for researchers to test programs in areas that are similar to conditions found in a wider population so that it is more plausible that results will generalize.

Another response to this criticism is to test whether in fact similar effects are found from implementing the same policy or program in different contexts in order to build up a theory and evidence base of when results generalize and when they do not. We cannot and should not test every approach in every geographic, cultural, or economic setting, but if we observe similar results emerging in different contexts, we can be more confident about the generalizability of the findings.

An increasing number of studies do this. A program providing assets, livelihoods training, and other services to the "ultra-poor" is being tested in seven different countries, and preliminary results suggest similar impacts (Innovations for Poverty Action 2013). A remedial education program was first tested in two cities in India with very different education systems (Banerjee et al. 2007). It was then tested in a rural context in India (Banerjee et al. 2010a) and is now being tested in Ghana. Again, very similar results have been found in all these contexts. Studies of basic microcredit products have found broadly similar results in urban India (Banerjee et al. 2010b), rural Morocco (Crépon et al. 2011), rural Mongolia (Attanasio et al. 2011), and the Philippines (Giné and Karlan 2011). In contrast, evaluations of similar programs designed to provide information to local communities and empower them to improve service delivery have produced very different results in different contexts (Glennerster and Kremer 2011). A program to encourage communities in Uganda to press for better health services was very effective

(Bjorkman and Svensson 2009), while a similar program in India to encourage communities to press for better government education services had no effect (Banerjee et al. 2010a). Whether it was the sector, the county, or the details of the program that made the difference is unclear.

Some studies seek to test more fundamental questions about human behavior on the assumption that these are more likely to generalize (although it is still necessary to test this assumption). For example, how does charging a small copayment for health prevention products change demand? Do people exhibit evidence of sophisticated procrastination and are they willing to pay for commitment devices? Randomized evaluations have found similar results on these questions in many different contexts, (Kremer and Holla 2009; Glennerster and Kremer 2011). This type of study arguably has wider social benefits because the results are less likely to be context-dependent.[7] [8]

b. Collection of Primary Data

Researchers working on randomized evaluations will, in most cases, collect primary data on individuals and/or communities. The collection and storage of these data places an ethical obligation on researchers to ensure that subjects provide informed consent for data to be collected and that the data will be treated confidentially. The collection and storage of primary data is far from unique to randomized evaluation, but complying with ethical standards represents one of the main areas where researchers working on RCTs interact with ethical regulations.

Economists have been complying with ethical standards in this area long before RCTs became popular, but important issues still arise (Alderman, Das, and Rao in this volume) especially as new forms of data are collected and data are collected in new ways. We discuss the practical ethical issues that arise for randomized and nonrandomized studies when collecting primary data in Section 4.

c. Close Collaboration with Antipoverty Programs

One of the biggest changes to how development economists work, which has come alongside the rise of randomized evaluations, is that researchers have become intimately involved in the design and implementation of antipoverty programs. The intensity of involvement of researchers varies considerably between studies. At one end of the spectrum, researchers may develop an idea for a new program (based on the results of previous research), design an evaluation to test the idea, and manage both the implementation of the program and the evaluation. At the other end of the spectrum, researchers can evaluate programs that others have designed and which have, for reasons other than evaluation, been implemented with an element of randomization. How should

[7] When testing a more general principle, we might want to combine evidence from lab-based randomized evaluations (which have the advantage of being less expensive to run) with field-based evaluations, which are more realistic and thus likely more reliable gauges of how people will respond to actual policies and programs.
[8] The creation of the American Economic Association's registry for RCTs (www.socialscienceregistry.org) is designed to make it easier to systematically assess the extent to which results generalize by making it easier to find all the studies conducted that test a given proposition.

ethical guidelines be applied across this spectrum of cases? Do they only apply to data collection or do they also apply to the program being implemented? The issue here is not randomization itself. Researchers using nonrandomized techniques may be involved in the design of the programs they evaluate. However, the increased use of randomized evaluations has been strongly associated with an increase in researchers' involvement in the design and implementation of programs as an integral part of their research.

Unfortunately, this is an area where the Belmont Report and subsequent guidance is less clear, at least as far as economic and social research is concerned. A distinction is made between practice (which is not governed by the regulations) and research (which is). The Belmont Report discussion is written from a medical perspective and a footnote explains 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." In the biomedical sphere the report says that even standard medical practice, when it is subject to systematic investigation, is covered by the guidelines. As we discuss below, if this were translated into economic research to mean that any program being systematically evaluated would itself have to follow research guidelines, this would prevent a lot of economic work going forward (this is presumably why the committee refused to extend this rule to economic and social work). Many other guidelines are also not very clear about the distinction between research and practice in the areas of economic and social research. Canada's Tri-Council Policy Statement has among the clearest definitions of research, while still being close in spirit to other definitions. It states, "For the purposes of this Policy, 'research' is defined as an undertaking intended to extend knowledge through a disciplined inquiry or systematic investigation."

In the world of development the lines become even more blurred because government and nongovernment organizations regularly conduct evaluations that are not considered to fall under research guidelines. Data are collected and analyzed by these organizations and conclusions drawn from the results but this is not considered research.

The Nature of Researcher Responsibility in Project Implementation

The collaboration between researchers and implementers raises a host of important questions about the ethical regulation of this interaction. Most fundamentally, if a researcher is involved in implementation, do the same ethical standards apply to them in their role as implementers (or partners in implementation) as apply to them in their role as researchers? At one level, it seems obvious when reading the Belmont Report that they do. In the canonical case of a medical trial, the risk of harm that is being balanced against the likelihood of benefit is often the risk that the vaccine or medicine being tested does harm. In other words, it is in large part the risk of harm from the implementation that the guidelines are designed to address. But it is just as clear that we cannot always hold researchers accountable for the ethical issues raised by the program they are evaluating.

Angrist et al. (1990) examined the impact of serving in the Vietnam War by using the draft lottery to identify individuals who were more and less likely to serve in the armed forces. If we took the Belmont principles for biomedical research coverage literally here, the program was being systematically investigated and thus the draft would fall under the guidelines. However,

from a practical standpoint, the researchers clearly had no ethical responsibility for the war and their IRB had no jurisdiction to require informed consent before men were drafted. Similarly, when the Government of Colombia held a randomized lottery to give vouchers to students to attend private school (Angrist et al. 2002) or the Indian Supreme Court decided that one-third of local village councils should be headed by women, and some states picked villages through a lottery (Chattopadhyay and Duflo 2004), research ethical guidelines did not apply to the implementation of the program. The randomization was done for reasons other than evaluation, such as fairness, and initially the agencies involved were not even aware that an evaluation was taking place. In these cases, the ethical guidelines only apply to the data collection efforts of the researchers.

But this distinction between cases where researchers are responsible for implementation and where they are not can be hard to make in practice. If the implementation was going to go ahead without the evaluation in any case, but the researchers work closely with the implementer on the logistics of implementation to enable an element of randomization, do the researchers take on responsibility for the ethics of the program? Does the entire program itself now come under review by the IRB? Or only those elements of the program that researchers modified in order to generate knowledge through systematic inquiry? What if in the process of designing the evaluation the researchers makes some minor suggestions which they hope will improve the program? What if the final decision about program design is with the implementer, and the implementer takes advice from many people (including the researchers)?

There are no simple answers to these questions, and different review boards take somewhat different positions. Some want to understand the risks and benefits of the program implementation even if it would have taken place without the researcher's involvement. Other review boards are mainly interested in the evaluation aspect of the research, especially if the program would go ahead in the absence of an evaluation.

One rationale to regulate implementation is that it is difficult to judge the level of researcher involvement and the degree of change in implementation resulting from research, so a more inclusive policy is preferable. Another, more cynical rationale is that a key objective of IRBs is to keep the name of their organization out of the newspaper, and even the association of one of their researchers with a program that has negative effects is bad publicity. Thus an IRB may want to stop evaluation of a risky program even if the program would go ahead without researcher involvement and there are large social benefits to understanding the risks.

In our view, if the definition of research is generating knowledge through systematic investigation, then only the elements of the program that were changed to allow for systematic investigation are research. However, because researchers may be tempted to define those changes narrowly, it is good practice to provide information on the whole program to the review board as well as setting out the counterfactual—i.e. what would have happened if the program was not being evaluated.

While the world of implementing poverty programs in resource poor settings is messy and complex and full of difficult tradeoffs that have ethical implications, the fact that researchers are now engaging with these issues, often quite directly, is in our view a very positive development.

The experience of directly implementing programs and/or working closely with those who do has influenced researchers understanding of and interest in the challenges of implementation in developing countries (Banerjee, 2007). And if we think research is a valuable endeavor and generates important lessons, then researcher involvement in these questions should be encouraged. The complex tradeoffs that implementers make would not disappear if researchers retreated from the field of policy implementation. Thus, as we consider how to regulate this nexus of research and implementation, we must be careful to avoid overregulation.

Potential Harm from Overregulating Researcher Involvement in Implementation

There are costs, and indeed even risk of harm, involved in applying research guidelines to implementation simply because a researcher is involved. In particular, IRBs often impose tighter consent requirements to participate in a study than are used to determine participation in most programs. This may be a consequence of the assumption that "randomized evaluation" means "medical trial" and therefore one-size-fits-all consent requirements are applied that do not take risk (or the lack of it) into account. A real example makes the point.

In a study on the educational impacts of mass school-based deworming in Kenya, the IRB overseeing the study decided at one point that written parental permission should be required before deworming drugs could be given to children. If the program had not been part of a study, this requirement would probably not have been in place given the WHO recommendation in support of mass school-based deworming and its conclusion that the risks are negligible. Acquiring written consent from parents is hard in Kenya, and the result was that some children in the treatment group did not receive deworming pills. Given the low level of risk, it could be argued that imposing written consent imposed harm and was not in line with ethical principles. Interestingly, this was not a case of US rules being imposed inappropriately on a developing country, as the requirement for written consent was imposed by the Kenyan review board.

Overly burdensome informed consent requirements could also make it difficult to work with a wide range of implementers and evaluate a wide range of important projects. Imagine a program designed to reduce child marriage. Such a strategy has risks: maybe if the family keeps a girl in school until the legal age of marriage, it will be harder to find a suitable partner for her. The implementing organization, however, does not want to start their program by warning about the risks of delaying marriage. Households are free to participate in the program and know the local marriage market well (certainly better than an IRB committee does). Stressing the risks at the outset of the program and documenting consent to continue from all households whether or not they participate in the program could undermine the partners' efforts to encourage families to delay marriage. Should the researcher refuse to work with such an organization?

Overregulation and peer pressure could also stop researchers from evaluating programs that are being carried out but might have a risk of harm. A naïve reading of the researchers' ethical obligation to avoid harm might be interpreted as meaning that researchers have an ethical obligation to ensure the program they are evaluating never does any harm to anyone. But we can never guarantee that no one will lose as a result of a program. If microcredit helps women start new businesses, existing businesses may be hurt. Even if we modify this criterion to say

researchers should only evaluate programs where most participants will gain in expectation, this is not, in our view, a correct reading of the Belmont principles.

Imagine there is a program that is quite commonly implemented but which a researcher is concerned may have harmful effects. The researcher does not have enough evidence to convince those running the program that it is harmful and should be shut down. Surely society would benefit if the researcher volunteers to evaluate the program and, if it has negative effects, help prevent more people from being hurt? In other words, it might well be that the benefits of gathering evidence of harm through a research program would outweigh the potential risks to those who take part in the research. In other words, such research could be compatible with the Belmont Principles.

The key criteria for moving forward, ethically, when a research believes that a program may be doing harm to the average participant would be: there is no hard evidence that the program is harmful (if there were we could use that evidence to have the program shut down without the evaluation); that all those taking part in the study are warned of the potential risks; and that the evidence generated is likely to be effective in reducing the prevalence of the program in the future if the results show it causes harm.

Some research may involve interventions that are socially desirable but have negative impacts on research subjects, such as programs intended to improve tax collection or regulatory enforcement. For example, McKenzie (2013) discusses an evaluation of government policies to get more large, informal firms registered for tax purposes. Should an IRB approve such a project? In our view, as long as the methods used do not themselves raise ethical concerns and the potential benefits to society plausibly outweigh the costs borne by the research subjects, it should.

Finally, it is important to make a distinction between cases where the researcher goes in thinking that a program may well be causing harm to the average participant and those cases where research finds that harm is being caused even though this was not anticipated at the outset. Criticizing researchers who happen to find the program they evaluate has negative impacts would create publication bias, which would be damaging to society as a whole.

Who Regulates Researchers Involved in Implementation?

Ethical regulation of research involving human subjects covers activities that are research, which are not always synonymous with involvement of researchers. For example, a researcher may know the evidence on what is an effective approach to addressing a particular problem and may advise governments to take up this approach or may even establish an organization to scale up the approach.[9] This work is not covered by human subjects regulation. A more complicated situation is where a researcher advises a government on the design of a program or sets up an NGO to implement programs initially with no research objectives. Later they decide that the program provides a good opportunity to test some important questions. Following the discussion

[9] Rachel Glennerster is on the board of Deworm the World, an organization devoted to scaling up school-based deworming programs. These activities, while they involve humans, are not research.

above, research would be defined as any changes in the program that were the result of the program being subject to systematic investigation. In our experience, however, the fact that researchers are involved in the design of the program (even for nonresearch purposes) or running the implementation makes it more likely that the program itself will be subject to research guidelines. This raises the question of why researchers involved in implementation should be regulated differently than others undertaking exactly the same activities.

The problem is that the process for regulation of implementation tends to be very different from those developed for research. Implementation regulation tends to be less codified, less transparent, and less consistent across countries. For example, if implementation is carried out by a government agency there are usually many layers of required approval, but there are rarely stated principles by which proposals are evaluated. On the other hand, in democracies at least these processes are potentially more accountable than IRBs. Similarly, most governments have procedures for regulating the activities of NGOs or other implementers with which researchers work. Should researchers only work in countries and sectors where there are well-functioning reviews of implementation agencies? Such a position would have many drawbacks: it would raise the difficult question of who decides whether the local review process is adequate, and it would prevent us learning about and improving policies in the least functional societies where there is often a strong need for policy improvement and evidence to support it.

While the principles of the Belmont Report probably represent a decent basis for judging the ethics of implementation, most of those reviewing development projects are not even aware of these principles. This can lead to jurisdictional conflicts as research review boards and implementation review processes disagree on the appropriate way forward. What if a government wants to implement a program that monitors attendance of teachers and does not want to give teachers a right to opt out of the program—do they have the right to impose this view against the view of an IRB committee in a foreign university? What if the government simply wants all teachers to be required to answer the survey, but is happy for the individual answer to be anonymized and thus ensure that no action is taken against individual teachers? One response is to say that the government can go ahead, but the researcher has to withdraw when there is disagreement between the two authorities. However, this fails to take into consideration the often strong leverage that comes with foreign researcher involvement: for example, funding may disappear if the foreign researcher leaves, putting the government in a difficult position.

d. When Researchers Change Who Receives a Program and How

One area where randomized evaluations are different from other research is the way the research methodology directly affects who receives services under the program and, sometimes, how these services are delivered. This is one of the clearest areas of researcher responsibility. Such research-generated changes in program allocation may occur in nonrandomized studies (see footnote 6 for an example). But randomized evaluations tend to have more implications than other methodologies for who receives benefits.

Critics of randomized evaluations have charged that allocating benefits randomly treats research subjects “merely as means to some end” (Ravallion 2012). In this view, having a methodology dictate who receives benefits might be seen as treating people as objects for experimental manipulation. If done only for the amusement of the researcher with no general benefit this

would violate the principles of Justice and Respect for Persons. But as discussed in Section 3a, randomized evaluations, if designed well, can generate wider benefits to society that need to be compared to any potential harm caused by changing the allocation of beneficiaries. There is nothing intrinsically unethical about allocation being determined in part on the basis of evaluation needs.

A more subtle objection is that random allocation of resources is a form of mistargeting (Barrett and Carter, forthcoming). In the absence of an experiment, implementers would likely want to provide the intervention to the individuals, households, or communities most in need. But in an experimental setting, even if researchers are careful to ensure that the entire sample population being allocated into treatment and control groups meets some criterion of "needy," there is always more information available on the ground that could help target the neediest of the needy. By allocating treatments randomly, the research in effect throws this information away.

The potential harm caused by changes in allocation of beneficiaries due to randomization will depend on the type of randomization methodology involved and the context. We therefore take the different randomization methodologies one by one.

Treatment lottery

Under the treatment lottery, a sample frame is chosen and units are randomly chosen from it to be offered access to the program being evaluated. Unlike some other methods of randomization, in a treatment lottery some participants in the study are never given access to the program. Is this ethical? The treatment lottery approach is often 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 target coffee farmers in Rwanda but only have 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. In assessing the potential harm from this approach, we have to ask: how would allocation of the program be decided in the absence of the evaluation? Would the program have tried to assess which were the two hundred neediest or most suitable farmers? Or would the program have decided to work in a district with roughly two hundred eligible farmers? Or accepted farmers on a first-come, first-served basis?

The potential for less optimal allocation of the program to arise as a result of evaluation is mainly an issue when program implementation is highly targeted and there is no ability to expand the geographic scope of the program. In our Rwanda example, if the program was planning to target the two hundred most suitable farmers in one district, it may be possible to instead target the most suitable farmers in two districts and then randomly pick two hundred of these for the program, without significantly diminishing the accuracy of the targeting of the program. There will be specific farmers who would have got the program if it had remained in one district who do not get it because of the evaluation, but this is not unethical because a) guaranteeing that no specific individual is worse off is not a required or feasible standard for judging ethics; instead the standard is whether coffee farmers in Rwanda in general are or are not likely to be negatively impacted by the research and b) we do not know whether the program will have a positive effect so we do not know that any individual farmer is worse off for not receiving the program. This geographic expansion with little change in targeting criteria is probably the most common way in which treatment lottery evaluations are introduced. When targeting is

already very precise and it is not possible simply to expand to a wider geographic area, using a lottery around the cutoff is one possible approach (see below).

Randomized phase-in

Under a randomized phase-in methodology, the final allocation of the program is usually kept exactly as it would have been in the absence of the evaluation, but the order in which different people or groups are phased in to the program is altered. This approach is often used where a phase-in is planned from the start. If the phase-in in the absence of the randomized evaluation were carefully designed to target the neediest, or those who would benefit most, early, then introducing a randomized phase-in will generate some costs which have to be weighed against the benefits of the evaluation. In practice, it is rarely the case that implementers determine the order of roll out based on a careful assessment of need. The order of phase-in is usually determined by logistical considerations—for example, the first people or communities to receive the program are those nearest the implementer's headquarters or nearest the road. By randomizing the order of the phase-in we create logistical complications for the implementer (which are an important cost of evaluation), but rarely does this approach lead to less needy communities jumping the queue over more needy communities.

Treatment lottery around a cutoff

Unlike a simple treatment lottery, this methodology explicitly recognizes that some potential participants may be more qualified than others and is used when programs have explicit criteria for ranking eligibility. The methodology takes potential participants who are near the cutoff for eligibility and randomly selects who will be accepted into 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. Is this ethical?

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.

Encouragement designs.

Under randomized encouragement designs, the evaluation does not alter who is eligible for the program: those who meet any program eligibility criteria in both treatment and comparison groups can take up the program. Instead, some randomly selected individuals or communities may be given more information about the program, or they may be given help in signing up for the program or a small incentive to sign up. A necessary condition for this type of evaluation is that the encouragement does not directly affect the outcome of interest (except through encouraging take-up of the program). If this criterion is satisfied, the encouragement approach allows researchers to estimate the impact of the program on those who are induced to take it up.

Encouragement designs ease some of the ethical and political concerns that may arise when randomized evaluations restrict or alter access to the program being evaluated. Everyone can have access to the program; the researcher simply makes it easier for some to have access than others. But this difference with other forms of randomization is really a matter of degree. If it becomes easier for some people to get access to the program, the researcher is still potentially (if the program works) giving a benefit to some over others. Thus the process of balancing the risks and benefits of the evaluation applies even in this case.

4. Common trade-offs and practical ethical questions

In this section we discuss the practical issues and trade-offs researchers face when undertaking randomized impact evaluations. As we have made clear, many of these issues are far from unique to those conducting randomized trials, but they are important and worth discussing. Any researcher collecting primary data on human subjects needs to consider issues of confidentiality and informed consent. Respect for persons means that in most cases we have to explain any risks and ask consent from those participating in the research. These requirements may seem obvious and simple, but researchers have to make difficult decisions about how to keep personal information confidential and how to ask for consent (also see Alderman, Das, and Rao in this volume). We also discuss when researchers may, for reasons essential to the social value of the research, not fully inform or even deceive respondents.

a. Confidentiality of primary data

Unless researchers get explicit permission from the subject, they have a duty to ensure that any private information collected on individual human subjects is kept confidential. (The exception to this rule is if data are already publically available). This means that only researchers involved in the study and those responsible for their oversight have access to information that could identify individuals, including names, addresses, telephone numbers, e-mail addresses, numbers related to government identification numbers or programs, account numbers, vehicle identifiers, full face photos, and any other information that could identify individuals. IRB protocols typically require that survey responses be numbered and that a code linking names and numbers be stored securely and separately from the survey data (in a separate locked file cabinet or in password-protected files on password-protected computers). The exact systems that will be used to ensure confidentiality have to be explained when the study is reviewed by an IRB.

When using paper surveys it is standard to have all identifiers on the first one or two pages of the survey. All sheets in the paper survey have a unique ID, and as soon as some basic checks have been performed by a field supervisor, the first page(s) containing identifying information are detached and stored separately from the rest of the survey, providing rapid anonymity. At some stages during the analysis it is usually necessary to match identifiers and the main answers to the survey, but this analysis is done in secure conditions.

Two developments have necessitated a change to this standard protocol. First, data are increasingly being collected electronically, and the equivalent to “ripping the front page off the

survey" had to be developed. Often identifiers and survey content can only be separated after the data are downloaded from the electronic data collection device. In some cases, this downloading takes place over the internet or phone system. However, to balance these challenges, most electronic data collection systems can be made password-protected so that access to identified data is restricted. Similarly, encryption software allows the data that are sent over the internet or phone system to be made confidential. Encryption software has also allowed researchers to share identified data more easily within the research group. Given that these systems are new, however, different protocols have developed for different institutions about the appropriate handling and sharing of electronic data.

The second technological development is the increasing use of geo-positioning data. Because geo-positioning data can be used to identify an individual it counts as confidential data, however, it is much more integral to analysis than, for example, a person's name. Thus a greater part of analysis now includes handling confidential data, making encryption and other safeguards much more important.

Researchers are also faced with a difficult dilemma when deciding how much geo-positioning data to include when they publish their data. Geo-positioning data allow researchers to study, for example, the spillover effects of an intervention on those who live or work nearby. It may therefore be necessary to have access to these data to replicate research. Economic datasets that include geo-positioning data also allow researchers to use existing data to answer new questions by linking different datasets together. For example, researchers can link rainfall data with economic data, or link economic data sets collected at different times to create community-level panels.

How should we balance the risk that publishing geo-positioning data will make it possible to identify individual subjects with the real benefit from being able to link different datasets together? We have to consider what is the likelihood that it will be possible to identify individuals if the data is made accurate to half a mile, 1 mile, 5 miles. We also need to consider the likelihood of harm to the subject if it is possible to identify them from the geo-positioning data. What benefits will be given up if the geo-positioning data is only published at a very high level of aggregation?

The answers depend crucially on the context of the particular study: how densely populated is an area, what other questions are included in the dataset, what proportion of people are being interviewed, and how accessible are the data to those who could potentially use it to harm participants. For example, it would not be appropriate to put on the web data with geo-positioning variables accurate to within 2 miles on HIV status of farmers in rural Montana. The low density of the population would make it relatively easy to identify individual subjects, neighbors would have access to the internet and could look up the data, and HIV status is a sensitive piece of information. At the other end of the spectrum, a survey that measured child height in rural Rwanda is unlikely to risk harm even if geo-positioning data are released down to 1 mile. Many children are likely to live within a 1 mile radius, most of those who would interact with the subjects would not be able to figure out which child was being referred to in a given survey even if they could access the data, and the people the child interacts with can observe their height anyway so the survey results are not revealing any private information.

Given the huge variation in risks, blanket rules about publishing geo-positioning data are unlikely to be optimal (see discussion in Section 5c). Unfortunately, fear of being accused of releasing confidential information is leading researchers to strip most useful geo-positioning data from datasets before publication.

b. Informed consent for data collection and exposure to experimental treatment

Usually when data are collected on research participants, human subject rules require that the participants give their informed consent. In other words, the researchers need to explain who they are and what the research is about and ask whether the respondent agrees to participate. The exceptions are if the data being collected are already public or if the risks are negligible and the burden of getting consent outweighs the benefits. Examples of data that are public are previously collected primary data, or noting down public behavior (such as whether people who walk down a public street are wearing a hat). Risks are often viewed as sufficiently minimal and informed consent is not usually required if personal identifiers are not collected. For example, we want to record how the weather affects people's mood. We stand on a street corner and ask people whether they are feeling happy or sad, both on cloudy and sunny days. If we do not collect responders' names, ages, addresses or telephone numbers, informed consent may not be required.

A more complicated question is what kinds of consent are required from research subjects. Some gray areas include the necessity of oral versus written consent, consent to provide information versus consent to be part of the program being evaluated, informed consent in clustered randomized evaluations, and occasions where some suspension of informed consent may be required for methodological reasons.

Written versus oral consent

Informed consent rules designed in developed countries are not always sufficiently adapted to local needs. For example, some IRBs require written consent from subjects to participate in any study, even if the risks are low and literacy rates are low. It is important to take the local context into account here. The costs of gaining written consent may be very high in areas with very low levels of literacy. Specifically, it may require finding a literate member of the community, having them sit down with the respondent to read the consent form, and asking the respondent to then make their mark on a special consent form. Similarly, Zywicki (2007) presents cases where IRB processes have made consent forms more technical and less readable, making it difficult for people with lower levels of education to understand them. In this volume Alderman, Das, and Rao discuss additional complications with written consent in developing-country contexts.

There is also a danger that overly stringent consent rules could lead to perverse results by preventing people from receiving services that could help them. This is one area where existing guidelines impose higher burdens on researchers than implementers would face in the absence of an evaluation, with the potential to cause real harm. An example comes from a study of a potentially life-saving medical treatment in an undisclosed country in Africa which was shut down because researchers were unable to secure signed consent forms in advance (Zywicki

2007) even though in the absence of the study signed consent would probably not have been required.

Suspension of informed consent for methodological reasons

The obligation to acquire informed consent can be waived if the risks are low and full knowledge by participants could lead to behavioral responses that undermine the usefulness of the research. The use of mystery clients is a good example. An enumerator may come into a store and ask the price of a good or service, attempt to register a crime, bargain for the price of a taxi ride, or interview for a job. In all these cases the purpose of the study is likely to be undermined if, prior to the interaction, the enumerator explains that the interaction is part of a study and asks for informed consent. In most of these cases the risk to the subject is small, although cases where bribes or illegal behavior is involved need to be regulated much more closely. Some authors have argued that suspension of informed consent has been given too readily (Ravallion 2012; Barrett and Carter, forthcoming).

Who is the subject of a randomized evaluation and thus from whom do we required informed consent?

Cluster randomized evaluations (those in which treatment is randomly assigned at the school, community, or some other level above the individual) pose further complications because they raise the question of who is the subject of the study. Is it only those individuals on whom data are collected or is it everyone in the cluster who receives the intervention?

McRae et al. (2011) argue that in certain types of medical cluster randomized evaluations, patients are not, in fact, research subjects and hence protections such as informed consent do not need to be applied at the patient level. These conditions occur in a relatively narrow set of interventions, such as training for healthcare providers, in which the patients are not directly intervened upon by investigators, do not interact with investigators, experience no manipulation of their environment that may compromise their interests, and do not provide identifiable personal information. In the case of training, the healthcare providers themselves are the research subjects, and they are responsible for deciding what is ethical to do for the patient, so the researchers themselves do not have ethical accountability for a change in the patient's treatment as a result of the intervention.

This discussion links to the previous one on whether the researcher is responsible for the program intervention or just the study. If we think the researcher is only responsible for the study, then informed consent is only required of those on whom we collect data. If the researcher is responsible for the program in general, and if we believe that ethical research standards apply to implementation, then we need informed consent from everyone affected by the program. In cluster trials, community-level consent is often sought from "gatekeepers" in political or administrative positions (for example, village heads) or through community meetings. In most program evaluations by economists when the program is available to all members of a community, individuals are still required to opt in. This opting in (after appropriate briefing from the implementer) can be considered a form of informed consent. We need to be much more careful when subjects cannot opt out of an intervention; for example, if the study involves adding chlorine to a municipal water supply (Hutton 2001).

When is informed consent not sufficient?

Even if a researcher intends to gain informed consent from participants, IRBs reserve the right to rule that a study should not go ahead. What justifies regulators denying an informed choice to individuals in this way? The assumption must be that the individual is unable to make a rational assessment of the risk. There are three main examples where this claim is made.

First, in the language of the Belmont Principles, participants may be potentially vulnerable. It may be hard to judge whether prisoners are fully free to decide on participation or are being coerced or feel they are being coerced. Similarly, children may not be able to judge risks effectively, and their parents may not always have their best interests at heart. Therefore, the IRB takes more responsibility for judging risks and benefits for these vulnerable populations.

Second, when the risks are very high a research regulator may not permit a study to go forward because in their view the risks outweigh the benefits. This view that "the regulator knows best" has been criticized in the US in recent years by patients' groups who argue that research regulators and medical regulators are too cautious about what drugs are tested and whether patients can try untested drugs. Patient advocates take the position that as the risks are borne by the patients, they should have the last word, and only they can judge the tradeoff between the risk of treatment versus the potential gain of a cure. Regulators point out that assessing risk is difficult and they have particular expertise in assessing the likely benefit of treatment. In most economic and social evaluations there is less need for technical knowledge to assess risks than in medical programs. Nevertheless it is worth keeping in mind that lab experiments suggest humans are not always good at assessing risks and that experts judge risks differently than laypeople (see Kahneman 2011, p. 137ff for discussion). Whether experts make better decisions about risk is not as clear.

The third argument for regulators to overrule an individual's choice is that an inducement to participate is so high that it undermines the ability of individuals to make a rational decision about risks and benefits. This argument is controversial because imposing preferences on others is likely to reduce welfare. Behavioral economics has, however, generated evidence that some people exhibit present bias and that people's choices are not always consistent over time. Some people even understand that they face this inconsistency and are willing to pay to restrict their own future choices. This inconsistency could provide support for regulation.

On the other hand, we need to be careful that we are not, as researchers, falling into a self-interested trap of our own. Imagine an industry which regulates itself and whose regulations state that it is unethical to pay those people whose time, ideas, and effort are key to the industry's success. If this were any other industry than our own, we might criticize such regulation as unethical cartel behavior designed to drive down costs. It certainly has an uncanny resemblance to regulations stating that universities should not pay athletes who generate such large revenue for them.

We therefore have offsetting considerations to keep in mind. Large upfront payouts may distort people's ability to judge risks, and taking up a lot of peoples' valuable time and giving no benefit is an imposition. With these considerations, how do we judge how much payment is too much?

As with so many of the issues discussed here, there is no simple answer. The amount will clearly depend on the wealth and income of the individuals we are working with. The prevailing daily wage for the average participant is often a good benchmark for judging whether an inducement is large. If a survey takes half a day to answer, we may give a monetary or nonmonetary gift equal to half a day's wage. More commonly, an inducement will be much less than that, almost a token gift. It may be culturally appropriate to give a small gift when coming to someone's house and spending considerable time with there. Alternatively, if people are close to indifferent about completing a survey, a small amount might tip the balance in favor of participation.

Another key consideration is what the inducement is designed to influence. Is it linked to participation in the survey or participation in the program? Is the inducement an integral part of the program, or is it designed to entice people to participate in a program which entails risks participants might otherwise be unwilling to take? Ethical issues mainly arise where we are using the inducement to offset a risk of harm from the program. Unlike in medical trials, inducements in economic and social studies are usually either given for participation for the survey and delinked from participation in the program or are the potential benefit of the program itself.

For example, an ongoing study in Sierra Leone by Rachel Glennerster and Tavneet Suri evaluated a program that introduced a new variety of rice seed. Farmers were randomly chosen to be surveyed and were offered an inducement of Maggi cubes (used to flavor stews) as a token of thanks for participating in a survey, which lasted roughly 2 hours. Half of the farmers were then randomly chosen to be offered the new rice seed at different prices ranging from free to full market price. Farmers were free to accept or reject the new rice seed independent of their participation in the survey. The seed was bred to have higher yield and shorter maturity but had no health risks. The risk was that yield would be lower than for traditional varieties because farmers did not know how to grow the new rice as effectively. Thus, there is a second inducement to participate, namely for some the benefit of receiving below market price seed. This latter is an integral part of the program and would take place in the absence of an evaluation.

It could be argued that the poor face no real choice about whether or not to participate in a program. They are so desperate they will accept any resources whatever the risk involved. But in fact, the poor take up some opportunities and turn down others. In the rice example above not all farmers took the subsidized rice, and a few did not accept the free rice (Banerjee and Duflo 2011 have a good discussion of this point).

Thus while there are reasons to worry about whether very high inducements to participate in a study may undermine people's ability to make rational choices, this is only really a concern when there are significant risks to participation which the inducement is trying to overcome. Even here we need to keep in mind the principle of respect for persons, which implies that we should make sure that people are fully informed of the risks and let them make the judgment of whether it is worth their while to participate.

c. Misleading Respondents

Another important and difficult question is when it is permissible for researchers to use tactics to elicit information from research subjects that they might not be willing to provide if asked directly, and to what extent deception is a legitimate approach to elicit this information. Historically, social scientists have argued against deception in laboratory experiments in order to protect the "public good" of subjects' trust. Different researchers may use a common pool of participants, e.g. students on a university campus, and the integrity of the study requires subjects to take instructions at face value. If participants are deceived once, this may change their behavior in future experiments, and Jamison, Karlan, and Schechter (2008) find some empirical support for this concern.

But the main ethical consideration is whether it is appropriate from the standpoint of the participants, not from the standpoint of future researchers. In the field, misleading respondents may be necessitated when researchers need to collect information about socially undesirable or illegal behavior. There are ways to address "social desirability bias," without outright deception. For example, we may ask respondent to evaluate a political speech. We may not tell the respondents that our aim is to judge attitudes to gender as this might influence how they respond to our questions. But we are not deceiving them: we say we want to learn about people's response to politicians and we do, we just don't highlight that we want to know about the response to the gender of the politician. The informed consent document will need to be approved, but there is little risk of harm and a good reason to not fully explain the full motivation of the research.

In other cases researchers have used more overt forms of deception. In a study by Bertrand and Mullainathan (2004) researchers randomly assigned African American-sounding or white-sounding names to the curriculum vitae (CVs) of fictional job candidates. The researchers submitted the CVs in response to published job openings in two US cities. They then tracked the difference in callbacks for interviews between apparently African-American and white candidates with otherwise identical CVs as a measure of racial discrimination.

Another form of deception is a use of trained actors in real-world settings. A study in Kenya and Uganda by Dizon-Ross and coauthors (2013) sought to determine whether health workers distribute subsidized health products to targeted recipients. The researchers sent actors into clinics to ask for subsidized products for which they were clearly ineligible (i.e. men asking for insecticide-treated bednets intended for pregnant women). The researchers obtained a waiver of the consent requirements for the clinic visits because the research design would have been compromised if the health workers knew they were being observed and the protections against identifying the clinics or workers were deemed adequate. In this volume, Alderman, Das, and Rao also discuss the use of "standardized patients," actors who are trained to portray standardized versions of medical cases.[10]

Is it ethical for researchers to mislead respondents in these ways? To answer this question researchers and IRBs must weigh the social value of the research and the risks of harm from misleading respondents. In all of the examples above there is an excellent case to be made for the social value of the research: gender and racial discrimination and poor delivery of health services

[10] Friedman (2011) provides a useful discussion of the ethical tradeoffs involved with this type of deception in light of the Belmont principles.

are serious social problems in the societies where these studies took place. In each case, the risks of harm from the research were small. For example, the costs imposed on firms in the discrimination study were the time required to read and respond to an additional application, which are likely small. In the Kenya and Uganda case the anonymity of the clinic staff who inappropriately sold nets was preserved and action was not taken against them. The greatest potential risk to human subjects may be from the researcher intentionally or unintentionally revealing information about the respondents, which is why confidentiality of information, as discussed above, is essential.

5. Gaps in and Concerns with the Existing Institutional Framework for Ethics Review

As we have seen, the ethics of experimental research in economics poses constant trade-offs among competing ethical claims and values. Many of these trade-offs come down to weighing the benefits of the knowledge generated by research against the potential or actual costs borne by research subjects, ranging from the opportunity cost of their time to the risk of being harmed by an intervention under study; and also against any violation of respect of persons or injustice that the research might entail. We have concluded that the Belmont Principles provide a good basis for judging these trade-offs. But this still leaves open the question of who should judge whether the benefits outweigh the risks in any particular case, whether the existing institutions designed to regulate these trade-offs are doing this well, and ways in which these institutions can be improved.

The primary responsibility for ethical research rests with researchers themselves. Yet researchers are human and are affected by the incentives they face. For example, psychological research shows that people's judgments are subject to "halo effects": a positive or negative evaluation in one dimension (e.g., the potential professional benefits to the researcher) tend to bias our perceptions in other dimensions (e.g., the ethics of a research project).[11] This tendency may cause even well-intentioned researchers to overweigh the benefits and underweigh the costs of a research project. Thus IRBs can be useful as an institutional safeguard by double-checking a researcher's judgment. In a similar way, most developing countries have some sort of review of development projects that review the implementation side of projects being evaluated. However, IRBs also have their own incentives and information constraints, which do not necessarily mean they will produce rulings or systems that are optimal for society.

In this section we discuss gaps and problems with the existing institutional review system for randomized evaluations carried out in developing countries. As Section 3 makes clear, many of the ethical issues arising in the context of randomized evaluations are the same as those that arise with nonrandomized evaluations. Thus most of the gaps and concerns raised here apply with equal force to nonrandomized evaluations.

[11] For a discussion of halo effects, see Kahneman (2011), p. 82ff.

a. Missing institutional structures

If we think that institutional review boards are useful for checking the trade-offs inherent in most research involving human subjects, then logically these institutions should exist wherever this type of research takes place. In fact, the coverage of IRBs for economic and social research is rather patchy. Most US universities with researchers undertaking social research in developing countries (whether randomized or not) have functioning IRBs to review research involving human subjects. However, many European universities that do not have medical schools do not have IRBs. In some cases research organizations do not have IRBs, but there are national medical IRBs. Some of these medical IRBs do cover nonmedical research, some explicitly do not, and many are ambiguous on the point. In a few of the poorest countries no review boards exist even for medical studies.

Over the last few years there has been considerable progress in establishing IRBs that explicitly deal with social and economic research involving human subjects. This move has primarily been motivated by researchers working on randomized evaluations of social programs seeking approval mechanisms for their studies. Thus, for example, both the Paris School of Economics and the Institute for Financial Management and Research (IFMR) in India established IRBs in 2009. As Alderman, Das, and Rao discuss in this volume, World Bank staff are actively discussing the creation of an ethical review board (currently the Bank relies of its member countries regulations and its internal review system for projects). That the increasing use of randomized evaluations should have spurred the creation of IRBs is somewhat surprising, as presumably many of these institutions collected data from human subjects long before randomized evaluations became popular.

An additional 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.

b. Lack of clarity on the mandate and coverage of existing institutions

In many countries it is unclear whether economic and social studies fall under the jurisdiction of medical research boards. This ambiguity causes confusion and stress on the part of researchers, although it may to some extent be deliberate.

What should the criteria be for an economics study falling under the jurisdiction of a medical research board? Some boards state that “health” studies fall under their jurisdiction. But what defines a health study? One in which any question about health is asked (that would include most economics and social science studies)? Or a health intervention is tested (note that a health intervention would include an SMS message encouraging people to get their child immunized)?

If a medical review board is the only IRB available in the country, one could argue that it is effective to build on this existing institution to provide coverage for economic and social research. However, it is important that if medical review boards extend their jurisdiction in this

way, they should bring on experts to help them assess economic and social research and adapt their procedures to take into account the different needs of this research.

In addition to medical research boards, many countries require that researchers get research permits before they begin a study. In most cases, "evaluations" are exempt from these requirements for approval of research. Why this seemingly arbitrary distinction between evaluation and research? The standard definition of research versus evaluation is that research generates general lessons, while evaluation is more directly related to a specific project. But why would a study that is sufficiently well designed that it produces general lessons generate more risk and thus need more regulation than a poorly identified evaluation? This is not just a developing country issue. Developed countries also restrict regulation to systematic study designed to generate general knowledge. The implication is that it's fine to harm subjects if your study is not systematic and you only hope to learn about your own program. A cynical answer is that countries understand that the process for acquiring research approvals are often so dysfunctional and delayed that requiring them for all evaluation work would prevent most organizations from undertaking most evaluations. This lack of evaluation would have negative consequences on the quality of projects in the country. It might also reduce the number of projects, as some implementers and donors might refuse to run or fund projects if evaluations were so hard to do.

An alternative explanation is that as most countries have a process by which NGOs and other development agencies report and receive approval for their activities (including their evaluation work), there is no need to require duplicate approval for these activities. The research approval process is designed to catch those activities that do not fall under this alternative reporting system. Thus, for example, a randomized evaluation of a project run by an NGO will be reported through the NGO activity approval process. But a survey conducted by a foreign researcher for a randomized or nonrandomized evaluation unrelated to an NGO project might need a research permit because otherwise the government would have no way of knowing of its existence.

In an ideal world, greater clarity about what work requires human subject approval and what falls under medical research boards would be useful. However, in the short run, pushing for clarity might lead medical research boards and research approval committees to define their jurisdiction as widely as possible. Given the limited capacity of many of these institutions to review economic and social studies, there is some danger to this approach. Indeed, it is quite likely that, knowing the limitations of these institutions, the lack of clarity is a deliberate strategy on the part of governments.

c. Overregulation and a reliance on rules versus discretion

Niskanen's (1971) model of bureaucratic behavior assumes that regulators and bureaucrats seek to maximize their budget. This raises the concern that once IRBs are established they will succumb to regulatory creep—expanding their regulation beyond the areas they were originally established to regulate. In addition to this concern, while IRBs are meant to weigh the risks and benefits of a study, their own incentives are likely to weigh the risks more heavily than the benefits. Those reviewing studies gain little of the upside if a study generates important knowledge for the world, while they face the downside risk of approving a study which ends up

causing harm. This is likely to make IRB reviewers overly cautious about what they approve, compared to what is socially optimal. Zywicki (2007) suggests that IRBs are overly sensitive to Type II errors (allowing potentially dangerous research to go forward) and insufficiently sensitive to Type I errors (incorrectly rejecting, delaying, or modifying a proposal because the board improperly overstated the risks to human subjects). Schrag (2011) documents many cases of what he describes as overreach by IRBs.

For an individual seeking to minimize effort and seeking to minimize the risk that they will be blamed for a bad decision, rules provide greater comfort than discretion. It is easier to justify a past decision by pointing to a rule than by explaining a judgment of costs and benefits as assessed at the time. But given that risks and benefits are likely to vary considerably between studies, rules risk undermining the Belmont Principles.

We discuss two examples of rules that have emerged in judging the ethics of studies involving human subjects. We argue that in most cases applying these rules to randomized or nonrandomized evaluations of economic and social programs in developing countries makes little sense.

Rules and HIPAA consent requirements

In 1996 the Health Insurance Portability and Accountability Act (HIPAA) introduced a set of privacy and security rules that regulate the storage and transfer of personally identifiable health data in the US. The regulations were designed to ensure the confidentiality of medical records that were increasingly being stored electronically. Procedures were established to gain consent for transferring health information from one user to another (for example, if a specialist was to transfer test results to a general practitioner). Anyone who has visited a doctor in the US will know that they will be asked to sign a very detailed HIPAA release forms, which are so long and dense that very few people read them.

But if HIPAA regulates the storage and transfer of medical data, and a research study involves collecting information on a person's health, and transferring this from the enumerator to the researcher, do subjects of the study have to sign a HIPAA release form? Some IRBs have concluded that they should. But these forms are arguably too dense to be effective in the US, let alone in countries with much lower levels of education. Arguably, having a page and a half of dense jargon read to someone is a less meaningful form of acquiring consent than a few clear sentences. Consent forms should be tailored to the particular study—how big are the risks that data will become publically available? How damaging would it be to the subjects if they did become publicly available? How can these risks be appropriately explained given the context?

Rules and geo-positioning data

A second area where, in our view, unhelpful guidelines have emerged is in the publication of geo-positioning data. The problem with geo-positioning data is that they can reveal the identity of a respondent even when other identifying information, such as name and address, have been stripped from the data prior to public release. As we argue in Section 4a, including geo-

positioning data in published data sets makes them much more valuable to other researchers. The risk that releasing geo-positioning data will enable an individual to be identified, as well as the potential harm caused if someone were identified, vary considerably according to the context and content of the dataset.

HIPAA rules (which apply to health data in the US) suggest two alternative approaches for ensuring that data have been appropriately "de-identified" prior to publication. The first approach is for a well-informed expert to determine if in isolation or in combination the variables to be released could identify an individual when combined with other available data. The second approach is to simply take out all variables on a pre-specified list (the safe harbor approach). Because this approach is designed to suit all data it has to veer on the side of caution. The safe harbor approach suggests that geo-positioning data should not be released at a level lower than a state. An exception is made for geographic areas formed by the first three digits of a zip code that contain more than 20,000 individuals. The problem arises when this safeharbor list is required (i.e. authors are not given the option of using the expert determination and is applied to data that are not as sensitive as medical records and are not from the US. We believe that applying this rule to these other situations is not in line with the Belmont Principles because it fails to take into account the different levels of risk of harm and benefits associated with different datasets. Applying the rule (without including the option of the expert determination) outside of the US also defies common sense. States are very different sizes in different countries. The state of Uttar Pradesh includes 200 million people; do we really think that prohibiting the publication of any geo-positioning data more precise that the entire state is needed to keep respondents responses confidential? If that were the case it would be a breach of confidentiality to say that a randomized evaluation of an education program took place in Jaunpur district in Uttar Pradesh.

d. Limited expertise in social sciences

Review boards that were established originally as medical review boards, or draw their members mainly from the medical sciences, often do not have enough expertise in the social sciences to properly judge the risks and benefits involved in a given study (Dingwall 2012). In particular, Dingwall argues that problems arise when boards fail to take into account the much lower risk involved in many social science studies than in many medical trials. What makes sense when regulating the provision of a new, untested drug may not make sense when testing the effect of doubling the number of government approved textbooks in a class.

Schrag (2011) argues that IRBs with medical backgrounds tend to apply standards from medical experiments without considering the differences between medical and social science research. Take the example of requiring written consent to take part in a study. If the study is designed to test a new vaccine whose risks are as yet unknown we may want to require written consent from participants even in a context with high levels of illiteracy where this involves finding a literate person the participant knows to carefully read the consent document to them and observe and validate their approval mark. However, oral consent may be sufficient when the risks of a study are minimal, as might be the case with an evaluation of a textbook distribution program. If a review body commonly addresses medical trials, it may have a blanket rule that written consent is always required.

There are also cases where IRBs impose rules from professional ethics documents even to those not covered by these rules. For example, there is a wide perception that US research rules require compensation to those harmed in the course of research and that treatment should be offered to those who are found to have medical issues. Doctors may have such responsibilities but these are not requirements under the common rule.
Medical reviewers may not appreciate the need for socioeconomic questions in a questionnaire. In one case, a review board rejected the inclusion of questions about assets in research about how regularly patients took their medication. The rationale was that asking about assets was too intrusive. From an economist's standpoint, questions about assets are standard. They are included to allow evaluators to test for heterogeneity of results by socioeconomic status: did the project work better for the rich than the poor? The same review board rejected the inclusion of questions about madrassa education. Again, in Muslim countries it is standard to distinguish between years of madrassa and non-madrassa education because of the very different subject matter covered in the different school systems. Given the program being evaluated required reading non-Arabic script and many madrassas only teach Arabic script, approval was given after appeal for this question to be included. However, this example highlights how differently economists and medical researchers think about these questions. Indeed, there are questions that medical and public health researchers routinely ask that social scientists might consider quite intrusive: about sexual activity or bowel movements. It is therefore important to have on review boards people who are experienced in running surveys of a given type and know what questions are sensitive and what are not.

It is worth noting that participants are always free not to respond to questions that make them uncomfortable. If we imposed a standard that no questions should be asked that might make someone uncomfortable, we would never have studies on important topics such as domestic violence.

e. Basic competency concerns

Any regulatory approval process creates rents and rents can generate rent-seeking behavior (Krueger 1974). Very large sums of money, years of preparatory work, and academic reputations are on the line when a study is considered for IRB approval. The risk of rent seeking behavior, which does not necessarily take the form of explicit requests for bribes, and how the risk can be moderated, should be taken into account when setting up regulations.

The major competency concerns that arise however are timeliness of approvals and inappropriate harsh requirements. Even if delays are simply an issue of bad management and not an attempt to extract rents from researchers, they nevertheless come with a high price. Zywicki (2007) notes that one of the costs of poor IRB regulation is that some research is never attempted for fear of the derailment by institutional review. We could add that some studies may not be attempted because of the required timeline for IRB approval would not allow researchers to act on time-sensitive opportunities. This is a particular issue for randomized evaluations where researchers have to coordinate their research with the timeline of implementers who may themselves be constrained by deadlines such as impending elections or agricultural planting seasons. The costs of IRB review due to "paths not taken" are difficult to observe and quantify.

One response is to establish competing IRB approval processes so than no one body has a monopoly on approvals and there are competitive pressures to provide good service to researchers (Zambia, for example, has two national approval bodies). The competing bodies must be composed of people with sufficient reputational capital to ensure that the competition does not lead to studies being approved when they should not be, but we are unaware of any cases where this has been a problem.

f. Capacity development and IRB

Some IRBs either explicitly or implicitly treat research studies involving local researchers more favorably arguing that this promotes capacity building. In our view because protection of human subjects and capacity building of local researchers are distinct objectives they should be treated separately and capacity building should not be a criteria by which to judge research proposals by IRBs. The sole concern of IRBs should be research subjects and those who might benefit from the research findings.

Conclusion

While there has been debate in other social science disciplines about the appropriateness of existing ethical guideline for social sciences, economists have not been very active in this debate until recently. The increased use of randomized evaluations of social programs in developing countries, however, has sparked more debate amongst economists about whether the existing systems of ethical oversight are adequate.

We conclude that many of the important ethical issues that arise when conducting randomized evaluations are far from unique to randomized evaluations. It is relatively rare that randomization itself raises difficult ethical issues (although we discuss examples where this may be the case). Alongside the use of randomized evaluations, however, has come a change in the way development economists work. Researchers have become much more involved in the details of program and policy implementation. Intense partnerships have developed between researchers and implementers that can blur the researcher/implementer distinction. It is possible that this change in role is in part responsible for the increased attention to ethical issues. It is certainly the case that many development economists are involved in designing projects that have direct consequences for the lives of thousands of people in developing countries. Whether we view this as a good or dangerous trend depends on whether we think that development economists are well equipped to advise on project design. One important but difficult question that arises from this trend is whether researchers who are designing programs should be regulated as researchers or as implementers. Current regulations are somewhat unclear with the result that programs influenced by researchers are often regulated both as research and as programs.

Finally, we have concluded that while the principles behind ethics regulations are reasonable and balanced, they are not always well implemented. In countries that have established medical review boards, simply extending these to cover social and economic research comes with risks. All too often IRBs apply rules that make sense for medical trials or are derived from medical professional ethics rules without taking into account the much lower risk involved in economic

and social studies. Real costs in research forgone are incurred when regulatory bodies are slow and overweigh risks.

References

Abadie, Alberto and Guido Imbens, forthcoming. "Estimation of the conditional variance in paired experiments." *Annales d'Economie et de Statistique*.

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, Erik Bloom, Elizabeth King, Michael Kremer and Juan Saavedra. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92(5): 1535–58.

Attanasio, Orazio, Britta Augsburg, Ralph de Haas, Emla Fitzsimons, and Heike Harmgart. 2011."Group lending or individual lending? Evidence from a randomised field experiment in Mongolia." IFS Working Papers W11/20, Institute for Fiscal Studies.

Banerjee, Abhijit. 2007. "Inside the Machine." In Banerjee et al., *Making Aid Work*. Boston Review.

Banerjee, Abhijit, and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York: Public Affairs.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Lindon. 2007. "Remedying Education: Evidence From Randomized Experiments in India." *The Quarterly Journal of Economics* 122(3): 1235-1264.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2010a. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India." *American Economic Journal: Economic Policy* 2(1): 1-30.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2010b. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." Working Paper, MIT, June 30.

Barios, Thomas, Rebecca Diamond, Guido Imbens, and Michal Kolesar. 2010. "Clustering, Spatial Correlations and Randomization Inference." NBER Working Paper No. 15760.

Barrett, Christopher B. and Michael R. Carter (forthcoming). "Chapter 7: Retreat from Radical Skepticism: Rebalancing Theory, Observational Data and Randomization in Development Economics." In *Field Experiments and their Critics,* D. Teele ed. New Haven: Yale University Press.

Bertrand, Marianne and Sendil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakesha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94(4): 991-1013.

Bjorkman, Martina, and Jakob Svensson. 2009. "Power to the People: Evidence From a Randomized Field Experiment on Community-Based Monitoring in Uganda." *The Quarterly Journal of Economics* 124(2): 735-69.

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.

Casey, Katherine, Rachel Glennerster, and Edward Miguel, 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan," The Quarterly Journal of Economics, Oxford University Press, vol. 127(4), pages 1755-1812.

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

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté. 2011. "Impact of Microcredit in Rural Areas of Morocco: Evidence from a Randomized Evaluation." Working Paper, École Polytechnique, March 31.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48: 424–455

Dingwall, Robert. "How did we ever get into this mess? The rise of ethical regulation in the social sciences." in Kevin Love (ed.) *Ethics in Social Research (Studies in Qualitative Methodology, Volume 12)*, Emerald Group Publishing Limited, 3-26.

Dizon-Ross, Rebecca, Pascaline Dupas and Jonathan Robinson. 2013. "Governance and Effectiveness of Public Health Subsidies". Mimeo.

Friedman, Jed. 2011. "Sometimes it is ethical to lie to your study subjects." *Development Impact Blog*, June 29. < http://blogs.worldbank.org/impactevaluations/sometimes-it-is-ethical-to-lie-to-your-study-subjects>

Glennerster, Rachel and Michael Kremer. 2011. "Small Changes, Big Results: Behavioral Economics at Work in Poor Countries." *Boston Review* March/April Issue.

Giné, Xavier, and Dean Karlan. 2011. "Group versus Individual Liability: Short and Long Term Evidence from Philippine Microcredit Lending Groups." Working Paper, Yale University.

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

Innovations for Poverty Action. 2013. "Ultra-Poor Graduation Pilots." < http://www.poverty-action.org/ultrapoor> Accessed 26 February 2013.

Jamison, Julian, Dean Karlan, and Laura Schechter. 2008. “To deceive or not to deceive: The effect of deception on behavior in future laboratory experiments.” *Journal of Economic Behavior & Organization* 68: 477-488.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* New York: FSG.

Karlan, Dean, and Jonathan Zinman. 2010. "Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts." *The Review of Financial Studies* 23(1): 433-464.

Kremer, Michael and Alaka Holla. 2009. “Pricing and Access: Lessons from Randomized Evaluations in Education and Health.” In *What Works in Development? Thinking Big and Thinking Small*, ed. Jessica Cohen and William Easterly, 91-119. Washington DC: Brookings Institution Press.

Krueger, Anne. 1974. “The Political Economy of the Rent-Seeking Society.” *The American Economic Review* 64(3): 291-303.

Levitt, Steven and John List. 2009. “Field experiments in economics: The past, the present, and the future.” *European Economic Review* 53:1–18.

McKenzie, David. 2013. “Doing Experiments with Socially Good but Privately Bad Treatments.” *Development Impact* blog, May 27. < http://blogs.worldbank.org/impactevaluations/doing-experiments-socially-good-privately-bad-treatments>

McRae, Andrew D., et al. 2011. “Who is the research subject in cluster randomized trials in health research?” *Trials* 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.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India.” *Journal of Political Economy*, 119(1): 39-77.

Niskanen, William A. 1974. *Bureaucracy and Public Economics.* Northampton, MA: Elgar.

Olken, Benjamin. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115(2): 200-249.

Ravallion, Martin, 2008. “Evaluation in the Practice of Development.” World Bank Policy Research Working Paper no. 4547.

Ravallion, Martin. 2012. “Fighting Poverty One Experiment at a Time: A Review Essay on Abhijit Banerjee and Esther Duflo, Poor Economics.” *Journal of Economic Literature.*

Schrag, Zachary. 2011. “The case against ethics review in the social sciences.” *Research Ethics* 7(4): 120-131.

U.S. Department of Health and Human Services (HHS). 1978. *Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research*. Available at http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

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