Decision Theoretic Approaches to Experiment Design and External Validity[*]

Abhijit Banerjee
Massachusetts Institute of
of Technology and NBER
banerjee@mit.edu
economics.mit.edu/faculty/banerjee

Sylvain Chassang
Princeton
University
chassang@princeton.edu
princeton.edu/∼chassang/

Erik Snowberg
California Institute
of Technology and NBER
snowberg@caltech.edu
hss.caltech.edu/∼snowberg/

July 15, 2016

Abstract

A modern, decision-theoretic framework can help clarify important practical questions of experimental design. Building on our recent work, this chapter begins by summarizing our framework for understanding the goals of experimenters, and applying this to re-randomization. We then use this framework to shed light on questions related to experimental registries, pre-analysis plans, and most importantly, external validity. Our framework implies that even when large samples can be collected, external decision-making remains inherently subjective. We embrace this conclusion, and argue that in order to improve external validity, experimental research needs to create a space for structured speculation.

JEL Classifications: C93, D70, D80
Keywords: randomization, self-selection, external validity, non-Bayesian decision making, ambiguity aversion

[*]Prepared for the *Handbook of Field Experiments*. We thank Esther Duflo for her leadership on the handbook, and for extensive comments on earlier drafts. Chassang and Snowberg gratefully acknowledge the support of NSF grant SES-1156154.

1 Introduction

1.1 Motivation

In the last couple decades, two of the most successful areas of economic research have been decision theory—and its close cousins, behavioral and experimental economics—and empirical microeconomics. Despite the fact that both emphasize experimentation as a method of investigation, there is almost no connection between the two literatures.[1] Indeed, there are good reasons why such a dialogue is difficult: an experiment designed according to the prescriptions of mainstream economic theory would get rejected by even the most benevolent referees; conversely, experimentation as it is practiced fails the standard axioms of subjective rationality.

Building on our work in Banerjee et al. (2014), this chapter seeks to establish such a dialogue. We believe that modern decision theory can provide a much needed framework for experiment design, at a time when experimenters seek to codify their practice. In turn, we believe that the issues facing the experimental community present a rich and useful set of challenges for decision theory. It is a rare opportunity for theorists to write models that could impact the practice of their colleagues down the hall.

1.2 Overview

We believe the main difficulty in finding a good theoretical framework for understanding experimental design stems from inconsistencies between the preferences of experimenters as individuals and as a group. As individuals, experimenters behave more or less like Bayesians. As a group however, experimenters behave like extremely ambiguity averse decision makers,

[1] See Chassang et al. (2012); Kasy (2013); Banerjee et al. (2014) for recent exceptions. This lack of connection despite the fact that economic theorists have extensively studied experimentation (Grossman and Stiglitz, 1980; Milgrom, 1981; Banerjee, 1992; Persico, 2000; Bergemann and Välimäki, 2002). Bandit problems have been a particular focus of this literature (Robbins, 1952; Bellman, 1956; Rothschild, 1974; Gittins, 1979; Aghion et al., 1991; Bergemann and Välimäki, 1996, 2006).

believing it is unwise to settle on a specific prior by which to evaluate new information.

Our framework considers the problem of a decision maker choosing both an experimental design and a decision rule—that is a mapping from experimental results into policy—who seeks to maximize her own subjective utility, while also satisfying an adversarial audience who may be able to veto her choices. We describe this framework, and then summarize the results in Banerjee et al. (2014): First, it unifies the Bayesian and frequentist perspectives. For small sample sizes, or if the decision maker places little weight on convincing her audience, optimal experimentation is deterministic and maximizes subjective utility. If instead the sample size is large, then randomized experiments allowing for prior-free inference become optimal. Second, the framework sheds light on the tradeoffs involved in re-randomization: It always improves the subjective value of experiments, but reduces the robustness of policy inferences. However, provided the number of re-randomizations is not terribly large (compared to the sample size), the robustness cost of re-randomization is negligible.

Having a model of experimenters also provides a useful perspective on pre-registration and pre-analysis. Bayesian decision makers do not need or desire either. On the other hand, a decision maker worried about an adversarial audience will value both. The important observation is that there is no need for the two perspectives to be seen as in opposition. Provided ex ante hypotheses are clearly labelled, there is no reason to constrain the dynamic updating of experiments as they are being run. Some decision makers will value knowing the ex ante hypotheses formulated by the experimenter, while Bayesian decision makers, who care only about the data collected, will value getting the most informative experiment possible. Reporting both, as "ex ante questions of interest," and "interim questions of interest" can satisfy both types.

The final sections are dedicated to the question of external validity. While there are ways to satisfy both the Bayesian and adversarial perspective in (policy) decision problems internal to the experimental environment, we argue that decision making in external environments is necessarily subjective—things may just be different in different circumstances. However, this does not mean that external inferences need to be vague or uninformative. We embrace the idea that external inference is necessarily speculative and that it should be thought of and reported as such as part of experimental research.

We formulate a framework for structured speculation that builds on two main observations. First, the manner of speculation, whether it is through a structural model or a reduced-form set of empirical predictions, is unimportant. What is important is for speculation to be stated as crisp hypotheses that can be falsified by further data. The advantage of structural modeling is that it automatically leads to a fully specified set of falsifiable predictions. However, model parameters are no less speculative than hypotheses formulated in natural language by experienced field researchers. While models have value in systematizing and clarifying thought, there is no formal reason to rule out any format of speculation experimenters are comfortable with, provided that predictions are made is a precise, falsifiable way.

The second observation is that creating space for structured speculation may have an important effect on how experiments are designed, run, and reported. Indeed, we believe it may result in a more effective and informative process of experimentation. We argue that the need for "better" speculation will lead experimenters to collect data that is ignored, unreported, or viewed as unconstructive to reported research: for instance, data on participant preferences and beliefs, the participants' place in a broader economic system, the role that values and norms play in the outcomes we measure, and so on. We illustrate this point by providing explicit examples of interesting topics for structured speculation.

The rest of this section very briefly discusses the history of experimental design, highlighting the interplay of theory and practice.

1.3 A Brief History

The first documented controlled experiment is found in the biblical book of Daniel, a story set around 605 B.C.E., comparing the health effects of a vegetarian diet with the Babylon court diet of meat and wine:

> Then Daniel asked the guard whom the palace master had appointed over Daniel, Hananiah, Mishael, and Azariah: "Please test your servants for ten days. Let us be given vegetables to eat and water to drink. You can then compare our appearance with the appearance of the young men who eat the royal rations, and deal with your servants according to what you observe." So he agreed to this proposal and tested them for ten days. At the end of ten days it was observed that they appeared better and fatter than all the young men who had been eating the royal rations. (Daniel 1:11–14, NRSV)

Despite the early emergence of controlled trials, it took millennia for randomization to be inserted into the process—by statistical theorists well versed in field experiments. Simpson and Pearson (1904) argues for a crude form of randomization in the testing of inoculants (while at the same time performing the first meta-analysis, see Egger et al., 2001) in order to establish a true control group. Over the years that followed, Pearson would formulate stronger and stronger defenses of randomization, emphasizing the need to draw controls from the same population as those that are treated (culminating in Maynard, 1909). Fisher (1926) was the first to provide a detailed program for randomization, which he expanded into his classic text on experimental design (Fisher, 1935).

Randomization became a mainstay of experimental design thanks to two factors. The first was medical practitioners looking for a way to evaluate treatments in a way that would prevent manipulation from the manufacturers of those treatments. Randomization alone proved insufficient to this task, which lead to the development of many tools, such as pre-registration and pre-analysis plans for trials, that we discuss in this chapter. These tools have had success in medicine, but their costs and benefits are likely to vary by field. As such, we have tried to identify, as abstractly as possible the factors that may make them more or less appealing, depending on the circumstances.

The second factor was a desire in many other fields of social science to identify the causal effects of interventions. Randomization was put at the center of frameworks for causal analysis leading, after some delay, to an explosion of randomized controlled field trials in several disciplines of the social sciences (Rubin, 1974; Pearl, 2000). Once again, however, randomization alone has not been sufficient to the task. Practical difficulties, such as treated participants being unwilling to receive treatment, have interfered. A number of statistical tools have been created to address these issues. However, as decision theory has little to say about the choice of statistical techniques, we do not discuss them here.

Finally, there is also work on experimental design that takes a Bayesian, rather than classical, perspective. However, like in econometrics, its presence is somewhat marginal. Even the proponents of Bayesian experimental design note that despite its strong normative appeal, it remains rarely, if ever, used (Chaloner and Verdinelli, 1995).

2 The Framework

We take the point of view of a decision maker who can inform her policy choice by running an experiment. She could be a scholar who is trying to come up with a policy recommendation, or a political entrepreneur trying to shape policy for the better. The decision problem can be *internal*, if the ultimate policy decision affects the population targeted by the experiment, or *external*, if it applies to a population different from that involved in the experiment (hence *external validity*).[2]

Our discussion and modeling follows Banerjee et al. (2014), but is more informal. The

[2]Note that the decision problem may differ because the population has changed—for example, it consists of different people, or the same people with different beliefs, or in a different context—or because the treatment differs in some way—for example, it is delivered at a different time, through a different distribution channel.

interested reader may consult the original paper for more details.

Actions and preferences. A decision maker needs to decide whether to implement some policy $a \in \{0, 1\}$, that provides a treatment $\tau \in \{0, 1\}$ to a unit mass population—which may be composed of people, districts, cities, schools, and so on—indexed by $i \in [0, 1]$ for *individuals*.[3] To inform her judgement, the decision maker is able to run experiments assigning a given number N of subjects to treatment or control.

Potential outcomes for subject i, given treatment τ, are denoted by $Y_i^\tau \in \{0, 1\}$. $Y = 1$ is referred to as a success. Each individual i is associated with covariates $x_i \in X$, where the set X is finite. Covariates $x \in X$ are observable and affect the distribution of outcomes Y. The distribution $q \in \Delta(X)$ of covariates in the population is known and has full support. Outcomes Y_i are i.i.d. conditional on covariates. The success probabilities, conditional on treatment τ and covariates x are denoted by $p_x^\tau \equiv \text{prob}(Y_i^\tau = 1 | x_i = x)$.

Environments and decision problems. To specify the decision problem, and the distinction between internal and external problems, we define environments z, which are described by the finite-dimensional vector p of success probabilities conditional on covariates and treatment status

$$p = (p_x^0, p_x^1)_{x \in X} \in ([0, 1]^2)^X \equiv \mathcal{P}.$$

For the first half of this chapter we consider internal decision problems in which the environment is the same in both the experimental and policy-relevant population. The second half puts more attention on external decision problems and external validity, in which the two environments may differ.

Given a known environment p and a policy decision $a \in \{0, 1\}$, the decision maker's

[3]For simplicity, we focus on policies that assign the same treatment status to all $i \in [0, 1]$.

payoff $u(a,p)$ can be written as

$$u(a,p) \equiv \mathbb{E}_p Y^a = \sum_{x \in X} q(x) p_x^a.$$

This formulation does not explicitly recognize unobservables, although it allows p_x^τ to vary in arbitrary ways as x varies, which is effectively the consequence of unobservables.

Experiments and decision rules. An experiment is a realized assignment of treatment to individuals represented by a tuple $e = (x_i, \tau_i)_{i \in \{1,\dots,N\}} \in (X \times \{0,1\})^N \equiv E$. Experiments generate outcome data $y = (y_i)_{i \in \{1,\dots,N\}} \in \{0,1\}^N \equiv \mathcal{Y}$, with each y_i an independent realization of $Y_i^{\tau_i}$ given (x_i, τ_i).

The decision maker's strategy consists of both a (possibly randomized) experimental design $\mathcal{E} \in \Delta(E)$ and a decision rule $\alpha : E \times \mathcal{Y} \to \Delta(\{0,1\})$ which maps experimental data—including the realized design e and outcomes y—to a policy decision a. We denote by $\mathcal{A}$ the set of possible decision rules. Since $\mathcal{E}$ is the set of possible probability distributions over the realized assignments of treatment, this framework allows for randomized experiments.

We assume that subjects are exchangeable conditional on covariates, so that experiments identical up to a permutation of labels are equivalent from the perspective of the experimenter (De Finetti, 1937).[4]

3 Perspectives on Experimental Design

3.1 Bayesian Experimentation

[4]The framework here is not particularly general. The goal is to provide us with just enough flexibility to illustrate specific issues. For example, we consider coarse policy decisions between treating the entire population or no one. In practice, one may consider more sophisticated policy decisions indexed on observable covariates. We also assume that the number of treatment and control observations are freely chosen under an aggregate constraint. In practice, the cost of treatment and control data points may differ. These simplifications do not affect our results.

Much of economic theory proceeds under the assumption that decision makers are subjective expected utility maximizers. As this implies Bayesian updating, we refer to such decision makers as *Bayesians*. While subjective expected utility maximization has been an incredibly useful framework, it leads to theoretical prescriptions at odds with experimental practice.[5]

Formally, let the decision maker start from a prior $h_0 \in \Delta(\mathcal{P})$ over treatment effects. In the context of our experimentation problem, optimal experiments $\mathcal{E}$ and decision rules α must solve,

$$\max_{\mathcal{E},\alpha} \mathbb{E}_{h_0}[u(\alpha(e,y),p)]. \tag{1}$$

An immediate implication of the subjective expected utility framework is that randomization is never strictly optimal, and for generic priors it is strictly sub-optimal.

Proposition 1 (Banerjee et al. (2014), Bayesians do not Randomize)**.** *Assume that the decision maker is Bayesian, that is, designs experiments according to* (1)*. Then, there exist deterministic solutions $e \in E$ to* (1)*. A mixed strategy (randomization) $\mathcal{E} \in \Delta(E)$ solves* (1) *if and only if for all $e \in \operatorname{supp} \mathcal{E}$, e solves* (1)*.*[6]

The intuition of the result is straightforward. Mixed strategies are never strictly optimal for subjective expected utility maximizers when a pure strategy equilibrium exists, and an RCT is a mixed strategy in the decision problem described above. Kasy (2013) uses a result similar to Proposition 1 to argue that randomized controlled trials are suboptimal. Specifically, it emphasizes that if the goal is to achieve balance between the treatment and control samples, this is more efficiently done by purposefully assigning participants to treatment and control based on their observables, so as to eliminate any chance of ending up with an unbalanced sample purely because of bad luck in the randomization process.

Proposition 1 is obviously at odds with experimental practice. Real-life experimenters go through non-trivial expense in order to assign treatment and control randomly. We interpret

[5]It is normatively appealing as well, and the "as if" axiomatization proposed by Savage (1954) seems so natural that subjective expected utility maximization is sometimes considered an expression of rationality.

[6]See Banerjee et al. (2014) for precise definitions and a proof.

this mismatch as an indication that the Bayesian paradigm provides a poor description of the objectives of actual experimenters. However, we also believe there is insight into experimental practice that can be gained by carefully considering Proposition 1. We do this in the following example, before turning to the adversarial perspective discussed in the introduction.

3.1.1 Example: The Logic of Bayesian Experimentation

Consider an experiment evaluating educational vouchers. This experiment will influence a school superintendent's decision of whether or not to introduce vouchers in her district. The superintendent has dismissed vouchers in the past, believing that by far the most important determinant of academic outcomes is whether a student is from a poor or privileged background. She has used this belief to explain the superior performance of private schools in her district, as they are a bastion for privileged students. However, in recent years, she has become open to the radical opposite of her belief: Schooling is the sole determinant of academic success. That is, even a poor student would do better at a private school. To test this hypothesis, she has convinced a private school to let her assign, however she likes, a single student to enroll there.

Faced with an experiment with a single observation, most academic experimenters would give up. How could anyone ever learn from such an experiment? What is the comparison group? Yet designing an informative experiment is easy: A Bayesian decision maker always has a prior, and she can compare the outcome of the child to that. Suppose the superintendent believes that a poor child can never score higher than the 70th percentile on a standardized test. She would then clearly find it informative if a poor child were given the lone spot in the private school, and then scored in the 90th percentile.

Adding a second child to the experiment brings new questions, and new insights. In particular, suppose that a slot in a public school is also allocated to this experiment. Should the child in the public school have an identical or different background to the student assigned to the private school? Should we allocate the private-school spot by lottery?

Once we recognize the role of the prior in setting the benchmark, these questions become easy to answer. Our superintendent starts from the theory that only background matters. Under that theory, the most surprising outcome, and therefore the one likely to move her prior the most, is one in which a poor child who goes to a private school significantly outperforms a privileged child who goes to a public school. If this occurs, she would strongly update towards the alternative explanation that schooling is all that matters. Thus, the optimal design involves giving the private school slot to a poor child and sending a privileged child to a public school. In particular, she is more likely to be impressed by the outcome of this experiment than one where both students are from the same background.

Strikingly, this example falsifies the idea that balanced treatment and control groups are intrinsically appealing. Moreover, we are arguing for a deterministic, rather than random, assignment of the students. Indeed, a lottery only moves us away from the ideal design: If the privileged child is assigned to the private school, very little can be learned.

Proposition 1 shows that this result applies for all sample sizes. The limits of this line of reasoning are only met if multiple decision makers with different priors (or a single decision maker unable to commit to a single prior) are involved. Introduce another school official with a slightly different prior beliefs about the effect of economic background: she believes that while a poor student would not benefit from a move to a private school, a privileged student would be harmed by moving to a public school. In this case the design suggested above is much less attractive. If we observe that the poor child does better, it could be either because the private school helps him to do better or because the public school hurts the richer child (or both!).

When the experimenter wants to convince other decision makers, she will design an experiment that not only informs her, but also informs members of her audience with arbitrary priors. This is the perspective that Banerjee et al. (2014) seeks to capture. In this setting, randomized experiments emerge as the only ones that successfully defend against all priors, that is, the only experiments whose interpretation cannot be challenged even by a devil's advocate.

3.2 Ambiguity, or an Audience

Although Bayesian decision-making is the default framework of economic theory, it is by no means a consensus. First, a decision maker may not trust her prior, exhibiting ambiguity aversion (Ellsberg, 1961; Schmeidler, 1989; Gilboa and Schmeidler, 1989; Klibanoff et al., 2005). Second, she may simply not be able to think through all possible implications of holding a particular prior, in effect violating Savage's completeness axiom (Gilboa et al., 2009; Bewley, 1998). Third she may recognize that she needs to convince others whose priors may diverge from her own.[7]

The model we propose in Banerjee et al. (2014) takes seriously the idea that experimenters care about convincing such an audience. This "audience" may actually reflect the experimenter's own self-doubts and internal critics, or a real audience of stakeholders with veto power (for example, referees).[8] The decision maker chooses the experimental design $\mathcal{E}$ and decision rule α that solve

$$\max_{\mathcal{E},\alpha} U(\mathcal{E},\alpha) \equiv \lambda \underbrace{\mathbb{E}_{h_0,\mathcal{E}}[u(\alpha(e,y),p)]}_{\text{subjective effectiveness}} + (1-\lambda) \underbrace{\min_{h \in H} \mathbb{E}_{h,\mathcal{E}}[u(\alpha(e,y),p)]}_{\text{robust effectiveness}} \tag{2}$$

where $\lambda \in [0,1]$. Here, h_0 is a fixed reference prior, while H is a convex set of alternative priors $h \in \Delta(P)$. A decision maker with these preferences can be interpreted as maximizing its usefulness under reference prior h_0, while also satisfying an adversarial audience with priors $h \in H$.[9] The first term captures a desire for informativeness from the point of view

[7] A related concern is that she may be accused of fraudulent manipulation of the evidence by those who disagree with her a priori. However, if outright fraud is a concern, verifiable procedures, more than randomization, become necessary.

[8] The model belongs to the class of maxmin preferences axiomatized in Gilboa and Schmeidler (1989).

[9] Note that if $\lambda = 1$, we recover (1), so that this model nests standard Bayesian expected utility maximization. If satisfying audience members was introduced as a hard constraint, then the weight ratio $\frac{1-\lambda}{\lambda}$ would be interpreted as an appropriate Lagrange multiplier for that constraint.

of the experimenter, and the second captures a desire for robustness.

Ambiguity Averse Experimentation. Banerjee et al. (2014) study optimal experimentation by ambiguity-averse decision makers under one additional assumption.

Assumption 1. *We assume that there exists $\nu > 0$ such that, for all $X_0 \subset X$ with $|X_0| \leq N/2$, there exists a prior $h \in \arg\min_{h \in H} \mathbb{E}_h(\max_{a \in \{0,1\}} p^a)$ such that for almost every $p_{X_0} \equiv (p_x^0, p_x^1)_{x \in X_0}$,*

$$\min \left\{ \mathbb{E}_h \left[\max_{a \in \{0,1\}} \overline{p}^a - \overline{p}^0 \middle| p_{X_0} \right], \mathbb{E}_h \left[\max_{a \in \{0,1\}} \overline{p}^a - \overline{p}^1 \middle| p_{X_0} \right] \right\} > \nu.$$

The condition says that even if an experiment were to reveal the probability of success at every value of the covariate x in X_0, there is still at least one prior in the set H under which the conditional likelihood of making a wrong policy decision is bounded away from zero.[10] We delay giving more intuition for this condition until after the following result:

Proposition 2. *For $\lambda \in (0,1)$:*

(i) Take sample size N as given. For generically every prior h_0, there exists $\underline{\lambda} \in (0,1)$ such that for all $\lambda \geq \underline{\lambda}$, the solution $\mathcal{E}^*$ to (2) is unique, deterministic, and Bayesian-optimal for $\lambda = 1$.

(ii) Take weight λ as given. There exists $\underline{N}$ such that for all $N \geq \underline{N}$, the optimal experiment $\mathcal{E}^*$ is randomized. *As N goes to infinity, the optimal experiment allows for correct policy decisions with probability going to one, uniformly over priors $h \in H$.*

Proposition 2 shows that the optimal experimental design depends on the number of available data points (or participants), and the weight the decision maker puts on her own

[10] The way this condition is specified implies that $N < 2|X|$.

prior versus those of the audience. Part (i) of the result shows that when sample points are scarce, or when the decision maker does not put much weight on satisfying anyone else (λ close to 1), optimal experimentation will be Bayesian. That is, the experimenter will focus on assigning treatment and control observations to the subjects from whom she expects to learn the most. Part (*ii*) shows that when sample points are plentiful and/or the decision maker cares about satisfying an adversarial audience, she will use randomized trials that allow for prior-free identification of correct policies.[11]

To build intuition for the result, and Assumption 1 it is useful to think of the decision maker as playing a zero-sum game against nature (with probably $1-\lambda$). After the decision maker picks an experiment, nature picks the prior which maximizes the chance of picking the wrong policy, given that experimental design. If there is any clear pattern in the decision maker's assignment of treatment, nature can exploit these due to Assumption 1. Randomization eliminates patterns for nature to exploit.

3.2.1 A Theory of Experimenters

Although randomization prevents nature from exploiting patterns in an experimental design, it is not always the optimal solution. There are two possible reasons for this. First, the decision maker may care so little about the audience (λ is close to 1), that preparing for the worst is of little use. Second, with small samples, the loss of power from randomization (relative to the optimal deterministic experiment) is so large that it offsets the benefit of reducing nature's ability to exploit a deterministic assignment. As the sample becomes large,

[11]Kasy (2013) reports a result that seems to directly contradict ours: that randomized experiments can never do better than a deterministic one, even with a maximin objective. The difference in the results comes from the fact that in Kasy's framework the audience sets its prior *after* randomization occurs, rather than between the revelation of the design and the actual randomization, as in our framework. In Kasy's framework, the audience will obviously pick a prior that means, in effect, that they can learn nothing from the actual treatment assignment.

Taking journal referees as an example of a skeptical audience, we believe our assumption is more realistic: Referees do show a fair amount of forbearance, even when faced with imbalance in covariates generated by a randomized control trial, though there are instances where they are sufficiently troubled by a particular case of imbalance to recommend rejection.

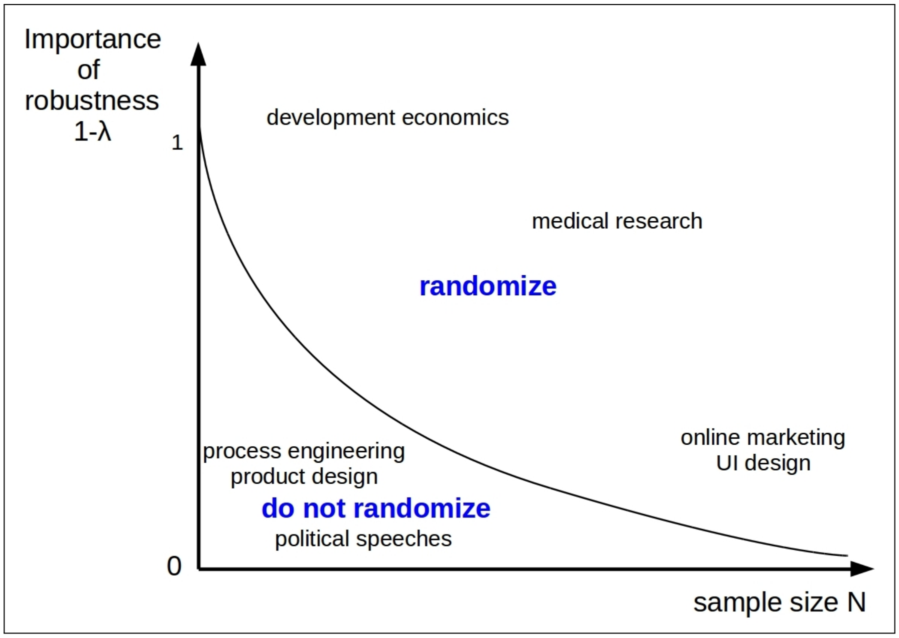


Figure 1: Different modes of experimentation

the loss of power from randomizing shrinks to nothing, while the gains from robustness against challenging priors remain positive and bounded away from zero.[12]

Figure 1 maps out implications of Proposition 2 for practical experiment design. In scientific research, when an experimenter faces a skeptical audience, she randomizes. In contrast, a firm implementing a costly new process in a handful of production sites will not try it on random teams. The firm will focus on a few teams where it can learn the most.[13] Yet when the available sample is large, firms do randomize. This is the case for firms dealing online with many end users: Although the firm only needs to convince itself of the effectiveness of a particular ad or UI design, observations are plentiful and randomization is cheap and used.

The logic of Proposition 2 applies at all stages of the decision-making tree that leads to the evaluation of a particular technology. When scientists want to convince others, they run

[12] As pointed out by Kasy (2013), the decision maker may also be able to limit the set of possible interpretations by deterministically choosing the right set of xs if there is enough continuity in p_x. Too much continuity is ruled out by Assumption 1.

[13] Similarly, a politician trying out platforms will do so at a few carefully chosen venues in front of carefully chosen audiences.

detailed randomized experiments. At earlier stages however, when a scientist decides what to experiment on, they do not just randomly pick a hypothesis. Instead, they develop a subjective prior on the technologies most likely to be worth exploring in detail. This fits well with our result: the number of experiments a scientist can run is limited, and each one of them is very costly, so it makes sense to subjectively refine the set under consideration. In online marketing, where experiments can be run at very little cost, there is much less need to use a subjective prior to refine the set of possible ads with which to experiment.

Additional implications of Proposition 2 refine our understanding of experimental practice. Part (ii) implies that a decision maker who randomizes even without understanding all its ramifications—why she is randomizing, what audience the experiment is meant to satisfy—will nevertheless produce an almost-optimal experiment for large values of N. Even if someone (or her own doubts) produces a particularly challenging prior, the decision rule is still likely to be close to optimal. In this sense, our approach addresses the concern that decision makers may violate Savage's completeness axiom.[14]

Proposition 2 also highlights the importance of actually randomizing. An experiment that adopts a protocol where assignment is only "nearly" random, such as assignment based on time of day of an experimental session (see Green and Tusicisny, 2012, for a critique), or the first letter of an experimental subject's name (as was the case in the deworming study of Miguel and Kremer, 2004; see Deaton, 2010 for a critique), will tend to find a skeptical prior in its audience. Randomization provides a defense against the most skeptical priors, but near-randomization offers no such protection.

4 Re-randomization, Registration, and Pre-Analysis

Proposition 2 suggests that the adversarial experimentation framework described by (2) may be useful for capturing the objectives of real-life experimenters. We now highlight ways

[14]A similar results hold for more complex policies that vary treatment with covariate x, provided the complexity of possible policies is limited. See Vapnik (1999) for operational definitions of "complexity".

in which having such a model can shed light on questions of current importance to the experimental community.

4.1 Re-randomization

Banerjee et al. (2014) brings the adversarial framework of (2) to bear on the question of re-randomization. A well-known problem with randomization is that it sometimes results in observable characteristics being poorly balanced across treatment and control groups (see Morgan and Rubin, 2012, and references therein).[15] Of course, stratification, blocking, and matching methods can be used to improve balance while maintaining randomization.[16] However, as any researcher who has tried to simultaneously stratify on multiple continuous variables knows, this can be quite difficult in practice. Moreover, these techniques have issues of their own (Athey and Imbens, forthcoming).

Re-randomization is a simple and intuitively attractive alternative: If a sample "looks" unbalanced, simply randomize again, and keep doing so until the sample looks balanced. While many authors caution against the use of re-randomization because it may have large statistical and internal validity costs (see Bruhn and McKenzie, 2009, and references therein), our framework can be used to precisely those costs.

From a purely Bayesian perspective, re-randomization does not create any concerns, and, indeed, may be beneficial, because it may select an experiment closer to the optimal deterministic experiment from a particular subjective point of view. That is, why should a Bayesian learn differently from the same balanced sample if it is reached by a single lucky

[15] Balance is important because it limits the possible set of alternative interpretations of the evidence, as described above. It also seems to serve as a rule of thumb for the experiment being competently executed, although this may not be warranted.

[16] Stratifying on several continuous variables is usually impractical for reasons related to the "curse of dimensionality". Consider an experiment with one treatment, one control, and four (continuous) variables describing participant heterogeneity. The natural strategy would be to bin subject characterizations along each dimension. In this example, we suppose each variable is split into five bins. Then there are $4^5 = 1,024$ cells in which to stratify, with each cell requiring two observations: one treatment, and one control. Unless the sample size is significantly greater than 2,048, with high likelihood there will be many cells with only one, unmatched, observation.

randomization, or by choosing among many?

In Banerjee et al. (2014), we show that the concerns brought up by Bruhn and McKenzie (2009) make sense in our adversarial framework. Re-randomization does have a cost in terms of robustness. Indeed, sufficiently many re-randomizations lead to an essentially deterministic allocation, which, we show, results in losses bounded away from zero for the adversarial audience. However, we also show that this cost is negligible, provided the number of re-randomizations is not exponential in the sample size.

We can make these costs and benefits precise: If K randomizations occur ($K = 1$ being a standard RCT), frequentist decision-making—that is, assigning the treatment that performs best empirically—is optimal up to a loss bounded by $\sqrt{\frac{max\{1,\log(K)\}}{N}}$. Importantly, $\sqrt{\log(N)}$ is a number between 1.5 and 3 for sample sizes between 10 and 10,000, which suggests that setting $K \leq N$ results in minimal losses of robustness. In turn, K randomizations guarantee that the final sample will be within the group of 5% most balanced samples with probability $1 - 0.95^K$. Observing that $1 - 0.95^{100} > 0.99$, this suggests the following rule of thumb for re-randomization.

Rule of Thumb:

Use the most balanced sample out of K randomizations, where $K = \min\{N, 100\}$.

Note that the balance criteria need not be defined ex ante. That is, the researcher can re-randomize K times, and select the assignment of treatment and control however they like *even after seeing the set of possible assignments.*[17]

We believe our proposal for re-randomization has several benefits. First, it provides simple, effective guidelines under which re-randomization is not problematic. Second, by doing so, it may help bring re-randomization out in the open. As discussed in Bruhn and

[17]Two important notes are in order here. First, when clustered randomization is done, for example, at the village level, then the number of re-randomizations should equal the number of clusters, not observations. Second, one can both stratify and re-randomize. That is, an experimenter can choose simple variables on which to stratify, and then re-randomize to achieve better balance on the more complex or continuous variables.

McKenzie (2009), many authors who employ re-randomization fail to disclose it, possibly because of the stigma attached to the practice. However, as long as re-randomization is done in a way that explicitly takes into account its costs and benefits, there is no reason for such a stigma.

Finally, re-randomization may help experimenters find compromises with governments or research partners uncomfortable with randomization. In some cases, experimenters negotiate a near-random treatment assignment scheme, as in the deworming example above. Our proposal is a middle ground: experimenters could produce a list of K randomizations to give to their implementation partner, and the partner could choose from that list. The criteria the implementing partner uses to choose a particular randomization could be anything they like it to be: from the one that "looks" the fairest to them, to more cynical ones that values having a particular village or person in the treatment group. Hybrids are possible as well: an experimenter could generate 100 randomization schemes, and allow the implementing partner to choose, however they want, from among the five most balanced.

4.2 Registration

Registration, enabled by platforms such as The American Economic Association's Randomized Controlled Trials Registry, is being embraced by a growing proportion of the experimental community. It has two effects. First, it creates a centralized and thorough database of experimental designs and outcomes that does not suffer from publication bias, file drawer bias, and so on. Second, it often leads researchers to commit to a particular experiment, and not change the experimental design during the course of the experiment. It should be noted that the latter is not a primary intention of registries, or their designers.

Within the framework described by (2), the first aspect of registration is unambiguously good. More information is always beneficial, simply because it can be ignored.[18]

[18]Note that decision makers exhibiting self-control problems (Gul and Pesendorfer, 2001), or decision makers with preferences over the revelation of uncertainty (Kreps and Porteus, 1978), may prefer to restrict

The commitment value of registration is much less obvious. In a dynamic setting, where experimental designs can be updated after the arrival of new information, Bayesians have no value for commitment, as they are time consistent. Indeed, if the decision maker is limited in her ability to specify complex contingency plans, then commitment has negative value. The value is even more negative when one considers the fact that updating a design may produce more useful information.

4.2.1 Good Commitment

Although registries are imperfect commitment devices, they are often used that way by experimenters. Commitment is valuable for the ambiguity-averse decision maker described by (2). Indeed, as Machina (1989) highlights, non-expected utility maximizers are not dynamically consistent. In other words, an ambiguity-averse decision maker who likes a particular design ex ante may be unsatisfied with the resulting experiment ex post, and try to alter its design.

The kind of temptation against which a decision maker may want to commit amounts to either: 1) tampering with realized random assignments, or 2) reneging from implementing a policy proven to be effective according to a burden of proof specified ex ante. Indeed, once a random assignment is drawn, there always exists priors under which the *realized* sample assignment and/or the policy conclusions are unsatisfactory. Commitment allows the experimenter follow through with the original plan. A plan, it should be remembered, that was ex ante satisfactory to both the experimenter and her adversarial audience.

The idea that registries allow various parties to commit to both an experiment and an action plan is plausible. Research partners may sometimes want to redraw assignments, shut down all or part of the experiment, or suppress parts of the data because they find the results misleading. Such hiding of information is likely to be undesirable in itself, in addition to its potentially harmful effects on the incentives of the experimenter (Aghion and Tirole, 1994).

the information available. Players involved in a strategic game may also have this preference.

Registration can reduce this risk.

4.2.2 Bad Commitment

There is, however, scope for excessive commitment. Indeed, while it is important for experimenters to commit to a randomized assignment, they need not commit to a specific treatment to guarantee robust inferences. For instance, after gaining experience with a treatment A, the experimenter may subjectively decide that a variant A' is likely to be much more useful. Experimenting with A' does not preclude robust inference about the value of A' versus the default alternative. In fact, data from experimenting with A and A' can be aggregated, corresponding to a mixture treatment A/A'.

In principle, if there are finite possible treatments, an ambiguity-averse decision maker may wish to randomize the treatments with which she experiments. In practice however, experimenters do not randomize the treatments they evaluate. The space of possible treatments is simply too large for such random exploration to be useful. Instead, the experimenter's subjective prior ends up driving the choice of intervention to evaluate. Randomized assignment after the treatment is chosen allows the experimenter to convince her audience to take the data seriously, although this may create a loss of valuable information.

If experimenters are bounded in terms of the number of possibilities they can imagine (as we definitely were), committing to a very detailed design once and for all makes little sense. It is costly to do, and it limits flexibility in ways that do not improve robustness. There is little reason not to update experiments, provided that these updates are registered, as is allowed (and tracked) by most registration platforms.[19]

[19] Of course, experimenters should not be allowed to first collect data, and then register a design that speaks only to a selected portion of this data.

4.2.3 Examples

An insider's perspective into Alatas et al. (2012) illustrates the cost of excessive commitment. Alatas et al. (2012) describes a field experiment in Indonesia where communities were asked to rank their members from poor to rich. The order in which households were considered for ranking was chosen randomly, driven by some concern for fairness. Households that were ranked earlier were ranked much more accurately, presumably because the rankers got tired or bored as the ranking meeting progressed. This was not something the authors had set out to learn. However, it might have made sense to change the protocol to guard against the inefficiency of late rankings—perhaps the ranking could have been done in batches, with breaks in between. But, the fact that the experiment was registered gave us a false sense that we were could not alter the design, even though such an update could have been reflected in the registry, and may have allowed for more learning.

Another example can be found in Andreoni et al. (2016), which used incentived choices to estimate time-discounting parameters of polio vaccinators in Pakistan. These parameters were then used to construct optimal contracts, which were tested against a standard piece-rate. Unfortunately, the authors had pre-registered, and thus felt committed to, a model of time preferences that the data showed to be mis-specified. This was a potentially fatal decision as the paper is a "proof of concept" of using preference estimation to design personalized contracts, and had the mis-specification been severe enough, it would have resulted in a failure to generate a significant improvement. Luckily, this was not the case, but it illustrates the dangers of "too much" commitment.

4.3 Pre-analysis Plans

A pre-analysis plan lists all outcomes of interest, and the ways in which data will be analyzed when the experiment is complete. Formally, it may be thought of as a subset of statistics S of the data.

4.3.1 Pre-analysis and Bounded Rationality

Interestingly, neither Bayesian nor ambiguity-averse decision makers find it beneficial to register a pre-analysis plan, nor would her audience care if she did. This follows from two implicit assumptions: 1) all data is disclosed, and 2) the decision maker and audience members have unbounded cognitive capacity. If an audience member is suspicious that the experimenter cherry-picked results, she can just run her own analyses. This seems appropriate when the experimenter faces a sophisticated professional audience (that is, referees, editors, seminar participants). However, in practice, there is demand for pre-analysis and thus, a careful, decision-theoretic foundation for pre-analysis plans is likely worthwhile.

While such a foundation is beyond the scope of this chapter, we can hint at a setup in which pre-analyis, that is, pre-selecting a subset S of statistics to be reported, becomes relevant. We believe this reflects the bounded rationality constraints of the decision maker or audience members. Indeed, if the decision maker can only process a subset of information S, she may be rightfully concerned about the way this set is selected. Formulating a pre-analysis plan can reassure the stakeholders, and facilitate actionable inference. Of course, if cognitive capacity is the issue, then pre-analysis plans cannot be excessively complicated: The goal is not for authors to anticipate all possible interesting inquiries into the data. This would defeat the purpose of pre-analysis plans by making them inaccessible to time-constrained decision makers.

In practice, experimenters are likely to speak to various audiences, each warranting different attitudes towards pre-analysis plans. A scholarly audience might reason that by demanding robustness checks, it is, in effect, forcing the reporting of all relevant dimensions of the data. Such an audience may prefer to ignore pre-analysis plans. However, an audience of time-constrained policymakers may behave differently, and only update from experiments with simple, clearly stated pre-analysis plans.

We see no need to view these perspectives as oppositional. Given the variety of audiences, the best response to us seems to allow for both ex ante and ex post analyses of the data within clearly defined "ex ante analysis" and "ex post analysis" sections. Ex ante-specified hypotheses will be useful to time-constrained audiences lacking the desire to really delve into the data. Ex post analysis of the data will allow experimenters to report insights that were hard to anticipate without the help of data.

4.3.2 Caveats

The discussion above does not touch on moral hazard concerns.[20] In this respect, two questions seem relevant: Is misbehavior by experimenters is prevalent in Economics? Are the mechanisms of registration and pre-analysis a long-term solution to this potential issue? The data at this point suggests that the answers are respectively "not very" and "maybe not". In particular, Brodeur et al. (forthcoming), find very little evidence of nefarious conduct in articles in top economics journals, and detect none in the reporting of results from randomized experiments. Moreover, in medicine, where norms of pre-registration and pre-analysis are often enforced by journals, a recent study by the Center for Evidence Based Medicine at Oxford University found that 58/67 of the articles examined contain misreporting—that is, failure to report pre-specified outcomes.[21] Response to these results has been quite varied, with at least one prestigious journal issuing corrections to all implicated articles, and another releasing an editorial defending aspects of misreporting.

A valuable aspect of pre-analysis plans that we do not account for is that they serve as contractual devices with research partners heavily invested in the outcome of an experiment (Casey et al., 2012; Olken, 2015). In these environments, a pre-analysis plan may prevent a research partner from shifting definitions after the data is collected. Additionally, specifying

[20] Humphreys et al. (2013) also emphasizes a communication role for pre-analysis plans. However, this should not detract from the very real commitment dimensions of registration and pre-analysis plans, and the fact that in order to make them successful, one needs to pay attention to how this commitment gives authors incentives to comply, or not.

[21] See http://compare-trials.org/blog/post-hoc-pre-specification-and-undeclared-separation-of-results-a-broken-record-in-the-making/.

table formats, and the analyses therein, ahead of time, is useful in identifying and eliminating disagreement between co-authors, and translating intentions into clear instructions for research assistants.

4.3.3 Theory

Pre-specified theories are potentially a way to protect against the accusation (and temptation) of motivated choices in analysis, while still preserving some analytical flexibility. This is in contrast with current common practice, which is to announce the theory in the same paper that shows the results of empirical analyses. Instead, a pre-specified theory should be "published" prior to empirical analysis, and, ideally prior to running the experiment.

An explicit, pre-specified theory preserves some flexibility of analysis, while restricting interpretation. It can be used to justify running regressions that were not pre-specified—those that are natural implications of the theory—without opening the door to full-scale specification searchers. Moreover, pre-specifying theory has the effect of making the experimenter's priors public, thereby allowing the audience to challenge an experimenter's interpretation of data on the grounds that it is inconsistent with the experimenter's prior theory. Of course some elements of an analysis cannot be derived from an abstract theory—for example, the level of clustering. Therefore it may make sense to combine some ex ante restrictions on specifications with a pre-specified theory.

It is worth emphasizing that even a theory that is developed ex post can impose useful restrictions on the analysis and reporting of results. In particular, given the implications of a theory, an audience can enquire why some are tested and others are not. While, in some cases, an ex post theory may turn out to be fully jury-rigged, it at least makes clear the author's assumptions, which then can be challenged.

In a sense, a pre-analysis plan is often just a reflection of an implicit theory, thus a pre-specified theory has some of the same drawbacks as a pre-analysis plans, but additional benefits. Like pre-analysis plans, theory will not change the beliefs of a skeptical audience that wants to examine the entire dataset. Additionally, if the theory turns out, ex post, to be irrelevant it may distract from useful features of the data (just as with pre-analysis plans). However, theory has the added benefit of making external extrapolation easier and more transparent, as we develop further in the next section.

5 External Validity

So far we have focused on internal decision problems, where treatment effects in the population enrolled in the experiment are the same as in the population a policy will be implemented upon. We now bring our framework to bear on external decision problems, in which treatment effects may differ between these two populations.

Formally, we allow the effectiveness of the treatment, described by vector p, to vary with the environment, denoted $z \in \{z_e, z_p\}$ (for experimental and policy-relevant):

$$p_z = (p^0_{x,z}, p^1_{x,z})_{x \in X} \in ([0,1]^2)^X \equiv \mathcal{P}.$$

While randomization is robustly optimal in internal decision problems ($z_e = z_p$)—provided the sample size is large enough—we now show that policy advice for external environments remains Bayesian even for arbitrarily large sample sizes. Under plausible assumptions, the best guess about which policy to choose in an environment or population that has not been studied is the experimenter's posterior after seeing experimental results in a related setting.

Let $H_{|p_z}$ denote the set of marginal distributions $h_{|p_z}$ over treatment effects p_z for priors $h \in H$ entertained by the audience. While information about environment z_e will likely affect the posterior over p_{z_p} for any given prior, it need not restrict the set of *possible* priors over p_{z_p}. This is captured by the following formal assumption.

Assumption 2. $H_{|p_{z_p}} \times H_{|p_{z_e}} \subset H$.[22]

External validity can be thought of as the following problem: after running an experiment in environment z_e, the experimenter is asked to make a recommendation for the external environment z_p. She thus chooses $\mathcal{E}$ and α to solve

$$\max_{\mathcal{E}\sim p_{z_e},\alpha} \left\{ \lambda \mathbb{E}_{h_e}[u(\alpha, p_{z_p})] + (1-\lambda) \min_{h \in H} \mathbb{E}_h[u(\alpha, p_{z_p})] \right\}. \tag{3}$$

Proposition 3 (external policy advice is Bayesian)**.** *The optimal recommendation rule α^* in (3) depends only on the experimenter's posterior belief $h_e(p_{z_p}|e)$ given experimental realization e. The optimal experiment $\mathcal{E}^*$ is Bayesian optimal under prior h_e.*

That is, external recommendations only reflect the beliefs held by the experimenter, not by the audience. This occurs because, under Assumption 2, evidence accumulated in environment z_e does not change the set of priors entertained by the audience in environment z_p—that is, it does not reduce the ambiguity in environment z_p. This further implies that the most information one can hope to obtain is the experimenter's subjective posterior belief over state p_{z_p}.

6 Structured Speculation

Proposition 3 formalizes the natural intuition that external policy advice is unavoidably subjective. This does not mean that it needs to be uninformed by experimental evidence, rather, judgement will unavoidably color it.

This also does not imply that subjective recommendations by experimenters cannot be used to inform policymakers. In many (most?) cases the policymaker will have to make a call without a randomized controlled trial tailored to the particular environment. As such,

[22]While this assumption is clearly stylized, our results generalize, provided there remains sufficient ambiguity about environment z_p, even conditional on knowing environment z_e very well.

the decision maker's most useful repository of information is likely to be the experimenter, because she is likely to deeply understand the experimental environment, previous results and evaluations, and how a policy environment may differ from experimental environments.

Proposition 3 also does not mean that external policy advice is cheap talk. Indeed, further evidence may be collected, and, provided that advice is precise, it may be proven to be right or wrong. What we should aim to do is extract the experimenter's honest beliefs about the efficacy of treatment in different environments. While this is not an entirely obvious exercise, we know from the literature on incentivizing experts that it is possible (see, for example, Olszewski and Peski, 2011; Chassang, 2013).

Practically, we do not think formal incentives are necessary to ensure truthful revelation. Instead, we believe a clear set of systematic guidelines for structured speculation may go a long way.

Guidelines for structured speculation:

1. Experimenters should systematically speculate about the external validity of their findings.

2. Such speculation should be clearly and cleanly separated from the rest of the paper; maybe in a section called "Speculation".

3. Speculation should be precise, and falsifiable.

The core requirements here are for speculative statements to be labeled as such, and be falsifiable. Practically, this means predictions need to be sufficiently precise that the experiment to validate or falsify them is unambiguous. This will allow testing by subsequent experimenters. By a reputational argument, this implies that speculative statements will not be cheap talk.

6.1 The Value of Structured Speculation

We believe that creating space for structured speculation is important and useful for several reasons.

First, providing a dedicated space for speculation will produce information that would not otherwise be transmitted. When assessing external questions, experimenters will bring to bear the full range of their practical knowledge built in the field. This includes an intuitive understanding of the mechanisms at work, of the underlying heterogeneity in treatment effects, how these correlates with observable characteristics, and so on.

Second, enforcing the format of speculative statements—that is, ensuring statements are precise and falsifiable—will facilitate and encourage follow-up tests, as well as interaction with closely related work.

Finally, to us, the most important side effect of asking experimenters to speculate about external validity is the creation of incentives to produce experimental designs that maximize the ability to address external questions. To address scalability, experimenters may structure local pilot studies for easy comparison with their main experiments. To identify the right sub-populations for generalizing to other environments, experimenters can identify ahead of time the characteristics of groups that can be generalized, and stratify on those. To extend the results to populations with a different distribution of unobserved characteristics, experimenters may elicit the former using the selective trial techniques discussed in Chassang et al. (2012), and run the experiment separately for each of the groups so identified.

While these benefits are speculative (and difficult to falsify!), it is our belief that creating a rigorous framework for external validity is an important step in completing an ecosystem for social science field experiments, and a complement to many other aspects of experimentation.

In the next subsections, we describe an operational framework for structured speculation *that can be used today.* We begin by providing concrete examples of what structured speculation may look like, and how it may be useful. We then propose a baseline set of external validity issues that should be systematically addressed. We conclude by discussing possible formats for structured speculation: qualitative, reduced-form, and structural.

6.2 Examples

To flesh out what we mean by structured speculation, we describe the form it may take in the context of a few papers.

Dupas (2014). Dupas (2014) studies the effect of short-term subsidies on long-run adoption, and reports that short-term subsidies had a significant impact on the adoption of a more effective and comfortable class of bed nets. In its Section 5, the paper provides an extraordinary discussion of external validity.

It first spells out a simple and transparent argument relating the effectiveness of short-run subsidies to: 1) the speed at which various forms of uncertainty are resolved; 2) the timing of user's costs and benefits. If the uncertainty over benefits is resolved quickly, short-run subsidies can have a long-term effect. If uncertainty over benefits is resolved slowly, and adoption costs are incurred early on, short-run subsidies are unlikely to have a long-term effect.

It then answers the question, "For what types of health products and contexts would we expect the same results to obtain?" It does so by classifying potential technologies into three categories based on how short-run (or one-time) subsidies would change adoption patterns:

Increased: cookstoves, water filters;

Unaffected: water disinfectant;

Decreased: deworming drugs.

While very simple, these statements are perfect examples of what structured speculation might look like. They attack a relevant policy question—the extension of one-time subsidies to other technologies—and make clear predictions that could be falsified through new experiments.

Banerjee et al. (2015a). This paper does not engage in speculation, but can illustrate the potential value of structured speculation for experimenters and their audiences. In particular, it reports on seven separate field trials, in seven different countries, of a program designed to help the ultra-poor. The basic intervention was the same in all countries, was funded out of the same pool, and the evaluations were all coordinated by Dean Karlan of Yale University.

Within the study, there were two options for external speculation. First, different countries were evaluated at different times. Second, there were multiple rounds of results for each location. Results from countries evaluated early in the experiment could have been used to speculate about the results from those evaluated later. Within a country, earlier rounds could have been used to speculate about later rounds. But what would have been the benefit of doing so? And how would we go about it, in hindsight?

There were many common questions that came up about this research: How long did we expect the effects to last? Was there any point in carrying out this program in rich or middle-income countries? Formally speculating about these questions in earlier rounds and countries would have provided a structure for answering those multiple queries, and justified elements of our experimental design that readers and reviewers had some reason to criticize. Additionally, making public predictions would have provided an opportunity for the authors—and other scholars—to learn about what kinds of predictions tend to be trustworthy.

Even *directional predictions*—a speculation that this effect will be larger than that one, or that it will be bigger or smaller than some number, possibly zero—would have been of some use. The point estimates of the program impact are smaller in richer countries. Does this mean the program needs to be re-thought for richer countries? We could have informed this decision by aggregating all we knew about the program, including the quantile results and results for certain sub-populations, to declare whether we believe the effects shrink with a country's GDP. We could have done this at different points in time, as the results came in from different countries and rounds, to see how good we are at making these sorts of predictions, and thus, how strongly we should advocate for our predictions at the end of the study. A similar exercise could have also been carried to predict the change in impact over time, which is key to understanding whether the intervention actually frees people from a poverty trap.

Banerjee et al. (2015b). Directional predictions would have also been very useful in maximizing the information from a series of so-called Teaching at the Right Level (TaRL) interventions described in Banerjee et al. (2015b). These interventions seek to teach children basic skills they lack, even when they are in a grade that presumes they have mastered those skills. This does not happen as a matter of course in most schools in the developing world (Banerjee and Duflo, 2011). As this intervention had already been shown to work on the margins of the school system, each experiment (RCT) focused on a different way of integrating this practice into government schools. The interventions varied from training teachers, to giving them the materials to use for TaRL, to implementing TaRL during the summer break (when teachers are not required to follow the set curriculum), to integrating TaRL directly into the curriculum, and so on. Each intervention built on the successes and failures of the previous interventions, culminating in two different, but successful, models. Yet without recording the predictions made along the way, this would look like an ex post rationalization of shooting in the dark. Even minimal public predictions—this approach is likely to work better than that—would have helped at lot.

Duflo et al. (2008). Another innovation that would have been useful is our call to record structured speculation at the end of each paper (in addition to in a repository, as we describe below). This would allow for a clear demarcation of results that are speculative—which will tend to arise in papers with pre-registration and pre-analysis plans—and those that are not. Such a demarcation would have clearly helped in dealing with Deaton's (2010, pp. 441–442) critique of the first TaRL paper (Banerjee et al., 2007).

> [W]hen two independent but identical [randomized controlled trials] in two cities in India find that children's scores improved less in Mumbai than in Vadodora, the authors state "this is likely related to the fact that over 80 percent of the children in Mumbai had already mastered the basic language skills the program was covering" (Duflo et al., 2008). It is not clear how "likely" is established here, and there is certainly no evidence that conforms to the "gold standard" that is seen as one of the central justifications for [randomized controlled trials]. For the same reason, repeated successful replications of a "what works" experiment, that is, one that is unrelated to some underlying or guiding mechanism, is both unlikely and unlikely to be persuasive.

Our proposal would have helped with such criticism by establishing a place within the paper where it was clear that this assertion was based on our own knowledge and intuitions, rather than a part of the experimental design.

6.2.1 A Post-hoc Evaluation

In summary, our proposal would have helped with criticisms of prior research in three ways. First, it would establish a place within research where such speculation is both expected and encouraged. Second, by attaching reputational incentives to such speculation, the reader can be assured that it is not just idle chatter intended to explain away an uncomfortable discrepancy. Third, because experimenters will be encouraged to speculate about the outcomes of replications before they happen, replications that are close to their predictions should increase, at least slightly, the credibility of the experimenter's preferred underlying mechanism.

An alternative approach, being pioneered by Stefano Della Vigna, of the University of California, and Devin Pope, of the University of Chicago, is to elicit priors on a specific research question from a wide range of experts (Della Vigna and Pope, 2016a,b). This has the benefit of forcing the audience to think about their priors before research is carried out, and identifying the places in which research can make the largest contribution by shifting the average prior, or collapsing the distribution of priors. However, it is unlikely to protect against the most skeptical members of an audience, who may not be in any given surveyed panel of experts. Moreover, it lacks many of the side benefits of our proposal.

On the other hand, the Della Vigna and Pope approach is being implemented today, while, with the exception of Dupas (2014), none of the papers above contained structured speculation. Why not? This was, of course, in part because it was not on the agenda. But there are deeper reasons: we, like many other researchers, focused on the reduced form LATE estimates, which tell us very little *directly* about how they would translate to other environments. A more natural basis for speculation would be to estimate a structural model, and use the estimated parameters—which can be made to directly depend on features of the environment—to predict out of sample. But, we must recognize that the choice of a model is itself subjective, so providing a model that rationalizes some prediction is not, in itself, completely reassuring.

However, the alternative may be worse. With different (Bayesian) readers having different priors and models of the world, even well-structured speculation without a model could be interpreted in multiple ways. The model serves as a currency for reducing, to a single number, the many disparate pieces of information that the author has. Without the prop of a model that exercise seems too hard to carry out with any accuracy.

To reduce the space of possible models, it would be helpful to demarcate the set of environments where structured speculation would be particularly useful, and the challenges likely to be encountered there. This is what the next subsection attempts to do.

7 Issues of Particular Interest

While our proposal could apply to any element of external validity, it is perhaps useful to outline a number of external validity issues that are focal for economists.

Focal External Validity Issues:

1. How scalable is the intervention?
2. What are treatment effects on a different population?
3. What are treatment effects on the same population in different circumstances?

Another important question that we do not discuss further is the one addressed by Dupas (2014): What is the effect of a different, but related, technology?

7.1 Scalability

A central concern in many development environments is how an intervention might *scale*—that is, how might the treatment effects measured in an experiment change if the intervention were rolled out across a province, country, or region? This concern is often composed of two inter-related issues: how spillover effects might enhance or reduce the benefits of a particular treatment, and how the incentives of an organization capable of implementing large-scale interventions might affect outcomes.

Spillovers. Spillovers encompass both general equilibrium effects and externalities. Consider an intervention that gives scholarships for top-performing students in local schools to attend provincial schools. As an experimental intervention, this policy may have large positive effects on a locality, because several students from the local school would be able to get an improved education. However, if rolled out nationally, the returns on human capital may diminish, possibly diminishing the treatment effect on outcomes such as wealth, savings, and consumption. There may, however, be positive general equilibrium effects. For instance, a more educated available workforce may increase FDI and lead to the creation of new types of jobs. General equilibrium effects are difficult to apprehend through purely experimental methods, but it is possible to draw on different sources of information to inform speculation. For instance, one may use regional heterogeneity in average human capital to map out what the effect may be should the program get rolled out.

Direct externalities (for example: physical interaction, information, contagion, and so on) may be more easily captured, as they tend to be more localized. Experimental designs that at least partially capture local externalities are now quite standard (Baird et al., 2014; Crépon et al., 2013; Muralidharan and Sundararaman, 2015). The difficult external validity question relates to the effect of scaling on adoption rates. In some cases, such as those of deworming drugs or vaccines, private returns are a diminishing function of aggregate adoption. This may be addressed by variation in the intensity of the program across locations. While this variation is likely to not result in sufficient power to become a main finding of a paper, it would be useful for guiding speculation about external validity.

Implementation by others. Informed speculation about implementing agencies is inherently difficult. Three environments seem relevant: implementations by other researchers, implementations by NGOs or international agencies, and implementations by provincial or country governments (Bold et al., 2013). The difficulty is that in order to make her speculation meaningful, the experimenter would need to specify the precise governments or NGOs that her projections apply to. This might expose the experimenter to political risk, and hamper her ability to conduct future experiments.[23] At the very least, it should be possible for the experimenter to highlight the specific aspects of the intervention that may make it

[23] This concern may be mitigated in practice by the fact that the employees of many organizations are more aware of their limited implementation abilities than researchers themselves.

difficult to be implemented by others.

One aspect of implementation that can possibly be controlled by experimenters is the reputational capital they have when they interact with the target population. They may be able to control for this by running initial perception surveys regarding their potential implementation partners, as well as by varying the way they present themselves. Having an official present at a meeting may significantly affect the way people engage with an experiment.

Again, an experiment may not be sufficiently well powered for variation in implementation to lead to significant findings. However, that data would clearly help informed speculation. In some cases, the experimenter may just have an intuitive understanding of how things would play out in different settings. Such intuitive understandings would be of great value to the next experimenter(s) who tried similar experiments. As such, it would be a useful contribution to speculate about the role of implementing agencies on outcomes.

7.2 Effect on Other Populations

If a program is effective in one region, or one country, is it effective in another? If a program is effective for a specific social group, would it be effective for different groups in the same country? For comparable groups in a different country? Answering such questions is inherently a speculative exercise, and yet it is useful for experimenters to do so. Experimenters have detailed intuitive knowledge of local mechanisms that can help clarify what matters for results to extend or not.

For example, suppose a program was found to be effective in India, and the experimenter tried to speculate about its effectiveness in Kenya. The experimenter may first assess the underlying heterogeneity in treatment effects and decide the program is principally effective in helping members of Scheduled Castes. If this is the case, one may reason that the program could be effective for historically discriminated populations of Kenya, say Muslims. However, by spelling this hypothesis out clearly, another experimenter may question its relevance if she believed affirmative action for Scheduled Castes appears essential for the treatment to be effective.

Subgroups and selective trials. We believe that subgroup analysis, which is often instructive but poorly identified, has an important role to play in formulating successful speculative hypotheses. Reweighting treatment effects by subgroups provides a natural way to project findings to different environments. This obviously includes groups formed on observable characteristics, say income, education, religion, and so on. Interestingly, this also includes unobservable characteristics elicited through mechanisms.

A recent strand of the experimental literature, illustrated by Ashraf et al. (2010); Berry et al. (2012); Cohen and Dupas (2010); Jack et al. (2013); Karlan and Zinman (2009), and formalized in Chassang et al. (2012), combines randomization with self-selection in order to "observe unobservables". The idea is as follows: randomized trials are lotteries over treatment. Many trials consist of a single lottery. By having multiple lotteries with different winning rates, and assigning costs to these lotteries, it becomes possible to elicit participants' values for the treatment, and estimate treatment effects conditional on values. This provides additional information helpful to project treatment effects on different populations.

For instance, as selective trials recover marginal treatment effects (MTEs, Heckman and Vytlacil, 2005), they allow the experimenter to figure out the effect of the program on populations selected through prices, by reduced availability, and so on. An experimenter will also make very different predictions about external treatment effects depending on whether it is effective for everybody, or only for highly-motivated participants.

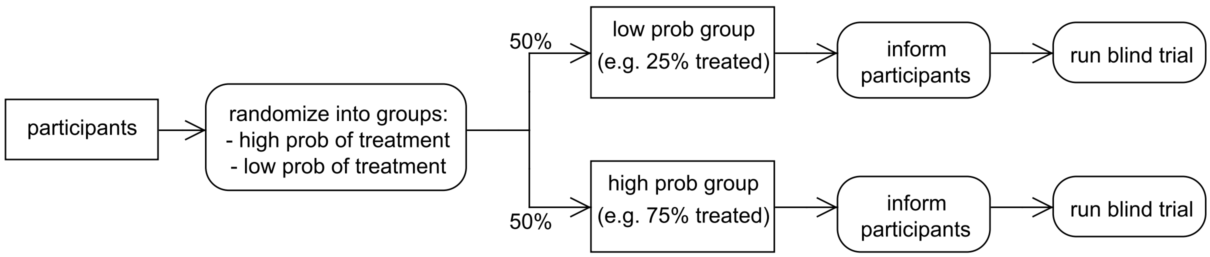
It is important to note that the "cost" of a lottery does not need to be monetary. Indeed, effort, more than money, seems to be a metric more easily comparable across locations. Alatas et al. (forthcoming) varies whether a participant has to travel for an interview, or can stay at home, to see if they qualify for a cash-transfer program for the poor. They find that those who travel are significantly more likely to actually be qualified for the program, and that interviewer coding of them is significantly more reliable. Randomizing the treatment (the cash-transfer program) conditional on whether or not the participant is judged to be qualified would have allowed this work to estimate returns for the motivated and for the less motivated. More generally, the variety of information that can be elicited through mechanisms is very large, and it frequently comes with a natural structural interpretation. We believe that collecting such information will prove helpful in formulating speculative hypotheses.

7.3 Same Population, Different Circumstances

The same population may react differently to treatment in different circumstances. For instance, if an intervention helps people save more, one may ask whether it will continue to be effective as people accumulate savings. Similarly, one may ask about the effectiveness of subsidies for technology adoption as information about the new technology spreads, and so on.

As before, subgroup analysis is likely to be helpful in forming opinions about the way effects will pan out as the population evolves. Richer participants or more informed communities may be used to proxy for future populations. As such, innovations in experimental design may also be helpful in this respect. Chassang et al. (2012) emphasizes that by either varying incentives, or by varying the participants' beliefs that they are being treated, it is possible to identify purely behavioral dimensions of treatment effects—that is, treatment effects that are due to participant's behavioral changes accompanying (the expectation of) treatment. For instance, a small business owner participating in an experiment in which she receives accounting training may decide to work longer hours because she believes the training is making her more productive. This could result in finding positive effects to train-

Figure 2: A Two-by-Two Blind Trial.



Notes: The figure shows the two stages of randomization, with participants first allocated to either a high- or low-probability treatment group, then informed of this probability (thus generating the corresponding placebo effect), and then receiving either treatment or non-treatment in a standard, blinded manner. Source: Chassang et al. (2015).

ing even when accounting itself is not useful. These effects, however, may not persist—the treatment effect for an informed participant, aware that accounting is of limited use, may be much smaller.

This observation is useful in medicine, where isolating treatment effects due to the interaction of a new drug with patient behavior is essential for understanding the true value add of that drug. In the context of medical trials, Chassang et al. (2015) proposes 2x2 blind trials (Figure 2) able to isolate both the pure effect of treatment, and the interaction of treatment and behavior. In a 2x2 trial, participants are randomly split into two arms. In one arm, the participants are told they will have a high probability of treatment, and in the other, they are told they will have a low probability of treatment. The trial within each arm is then run accordingly. Under the assumption that participant behavior will change with the probability of treatment, the trial independently randomizes both behavior and treatment. This is sufficient to isolate the pure effect of treatment, the pure effect of behavior, and the interaction of treatment and behavior.

Bulte et al. (2014) implements a version of a 2x2 trial in a development context using seed varieties as its technology. Its findings suggest that purely behavioral responses—that is, mediated by expectations of change—to treatment are significant. This, in turn, suggests that participants may change their response over time, as they learn the true effectiveness of a treatment. Practically, running the complex mechanisms described in Chassang et al. (2012), or using blind treatments as in Bulte et al. (2014) may not always be feasible. However, it should always be possible to survey participants about their expectations for the technology, and about how they are changing their practice in response to treatment. These survey measures, which in many cases would not naturally be reported as hard evidence, may prove quite useful in shaping speculative hypotheses.

7.4 Formats for Structured Speculation

We conclude our discussion of structured speculation with a brief discussion of the formats in which structured speculation may be expressed. We argue that, at this stage, there is no wrong way to formulate hypotheses about external validity, *provided the hypotheses are formulated in a clear and falsifiable way.*

Qualitative predictions. Simple qualitative statements are not necessarily less rigorous than analytical ones. For instance, Dupas (2014), describes environments in which treatment effects are likely to larger or smaller. These descriptions are simple, yet precise and falsifiable

Experimenters ofter produce reduced form estimates for multiple sub-populations and/or quantile treatment effects—though they may not report all of them—and these, along with some intuitive understanding of which environments are similar, make it possible for them to predict the direction of change in the treatment effect, though perhaps not the magnitude of the change. Such speculation is naturally expressed qualitatively, as in, "This treatment effect is likely to be larger than that."

Predictive models. If sufficient data about subgroups is available, experimenters may feel comfortable producing statistical models predicting treatment effects in other environments conditional on observables. Multi-period, multi-country trials such as Alatas et al. (2012), or multi-trial meta-analyses offer natural starting points. The main advantage of producing a fully-specified, predictive model is that it is unambigous, and by construction, clear and falsifiable. It is therefore a better starting point for further analysis than purely qualitative predictions. Note that the model need not necessarily predict point estimates. A model predicting ranges of estimates would be equally well specified.

Theory and structural models. Theory has an important role to play in formulating useful speculative hypotheses. If an experimenter lays out a theoretical model she thinks best summarizes the facts she sees, and the theory is rich enough to cover environments beyond what is in her experiment, she is effectively making a directional prediction for other environments.[24]

This is already happening to some degree. For example, Karlan et al. (2012) evaluates two interventions to improve business for tailors in Ghana. In one intervention, the tailors were provided with a cash grant; in the other, they were given training. Both changed the way the tailors practiced business (at least briefly), but neither increased profits over the long term. The authors develop a model in which this occurs because tailors treat these interventions as opportunities to explore new opportunities. While most of these fail, the option value of experimentation is likely still positive. This implies that there should be some tailors who experience very large gains from these interventions, but that, on average, the effect will be small and difficult to detect. To test this prediction, the authors look at other studies of similar interventions that are powered to detect differential changes in the right tail of the distribution. They find some support for their theory.

[24]Some predictions may be ambiguous, which is both a benefit and a drawback of formal models.

Identified structural models, just as predictive statistical models, are attractive because they make fully-specified predictions in external environments. An advantage they have over purely statistical models is that they can make the process of external extrapolation more transparent. We emphasize, however, the cautionary implications of Proposition 3. In all external decision making problems, inference is unavoidably subjective. In structural modeling, the source of subjectivity is the model itself.

8 Conclusion

This chapter has ranged from models of experimentation, to prescriptions for experimental design, all the way to external validity. Hopefully the wide range of (potential) applications of decision theory to experimental practice is enough to convince theorists and practitioners alike that this is a fruitful area for further discovery.

References

Aghion, Philippe and Jean Tirole, "The Management of Innovation," *The Quarterly Journal of Economics*, 1994, *109* (4), 1185–1209.

_ , **Patrick Bolton, Christopher Harris, and Bruno Jullien**, "Optimal Learning by Experimentation," *The Review of Economic Studies*, 1991, *58* (4), 621–654.

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias, "Targeting the Poor: Evidence from a Field Experiment in Indonesia," *American Economic Review*, 2012, *102* (4), 1206–1240.

_ , _ , _ , **Benjamin Olken, Ririn Purnamasari, and Matthew Wai-Poi**, "Self-Targeting: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, forthcoming.

Andreoni, James, Michael Callen, Yasir Khan, Karrar Jaffar, and Charles Sprenger, "Using Preference Estimates to Customize Incentives: An Application to Polio Vaccination Drives in Pakistan," 2016. NBER Working Paper Series # 22019.

Ashraf, Nava, James Berry, and Jesse M. Shapiro, "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia," *American Economic Review*, December 2010, *100* (6), 2383–2413.

Athey, Susan and Guido W. Imbens, "The Econometrics of Randomized Experiments," in Esther Duflo and Abhijit Banerjee, eds., *Handbook of Field Experiments*, Elsevier, forthcoming.

Baird, Sarah, J. Aislinn Bohren, Craig McIntosh, and Berk Özler, "Designing Experiments to Measure Spillover Effects," 2014. PIER Working Paper #14-032.

Banerjee, Abhijit, "A Simple Model of Herd Behavior," *The Quarterly Journal of Economics*, August 1992, *107* (3), 797–817.

_ **and Esther Duflo**, *Poor Economics: A Radical Rethinking of the way to Fight Global Poverty*, PublicAffairs, 2011.

_ , _ , **Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry**, "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries," *Science*, 2015, *348* (6236), 1260799.

_ , **Rukmini Banerji, James Berry, Esther Duflo, H Kannan, S Mukherji, and Michael Walton**, "Teaching at the Right Level: Evidence from Randomized Evaluations in India," *MIT Working paper*, 2015.

— , **Shawn Cole, Esther Duflo, and Leigh Linden**, "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics*, August 2007, *122* (3), 1236–1264.

— , **Sylvain Chassang, Sergio Montero, and Erik Snowberg**, "A Theory of Experimenters," 2014. Princeton University, *mimeo.*

Bellman, Richard, "A Problem in the Sequential Design of Experiments," *Sankhyā: The Indian Journal of Statistics*, April 1956, *16* (3/4), 221–229.

Bergemann, Dirk and Juuso Välimäki, "Learning and Strategic Pricing," *Econometrica*, September 1996, *64* (5), 1125–1149.

— **and** — , "Information Acquisition and Efficient Mechanism Design," *Econometrica*, 2002, *70* (3), 1007–1033.

— **and** — , "Bandit Problems," 2006. Cowles Foundation discussion paper.

Berry, James, Greg Fischer, and Raymond Guiteras, "Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana," 2012. Cornell University, *mimeo.*

Bewley, Truman F, "Knightian Uncertainty," in Donald P. Jacobs, Ehud Kalai, and Morton I. Kamien, eds., *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures 1983–1997*, Cambridge University Press: Econometric Society Monographs, 1998, pp. 71–81.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, "Scaling up What Works: Experimental Evidence on External Validity in Kenyan Education," 2013. Center for Global Development Working Paper #321.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, forthcoming.

Bruhn, Miriam and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 200–232.

Bulte, Erwin, Gonne Beekman, Salvatore Di Falco, Joseph Hella, and Pan Lei, "Behavioral Responses and the Impact of New Agricultural Technologies: Evidence from a Double-Blind Field Experiment in Tanzania," *American Journal of Agricultural Economics*, 2014, *96* (3), 813–830.

Casey, Katherine, Rachel Glennerster, and Edward Miguel, "Reshaping Institutions: Evidence on Aid Impacts using a Preanalysis Plan," *The Quarterly Journal of Economics*, 2012, *127* (4), 1755–1812.

Chaloner, Kathryn and Isabella Verdinelli, "Bayesian Experimental design: A Review," *Statistical Science*, August 1995, *10* (3), 273–304.

Chassang, Sylvain, "Calibrated Incentive Contracts," *Econometrica*, 2013, *81* (5), 1935–1971.

_ , Erik Snowberg, Ben Seymour, and Cayley Bowles, "Accounting for Behavior in Treatment Effects: New Applications for Blind Trials," *PLOS ONE*, 2015, *10* (6), e0127227.

_ , Gerard Padró i Miquel, and Erik Snowberg, "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, June 2012, *102* (4), 1279—1309.

Cohen, Jessica and Pascaline Dupas, "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics*, 2010, *125* (1), 1–45.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment," *Quarterly Journal of Economics*, 2013, *128* (2), 531–580.

De Finetti, Bruno, "La Prévision: Ses lois Logiques, ses Sources Subjectives," *Annales de l'institut Henri Poincaré*, 1937, *7* (1), 1–68.

Deaton, Angus, "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, June 2010, *48* (2), 424–455.

Della Vigna, Stefano and Devin Pope, "Run This Treament, Not That: What Experts Know," 2016. University of California, *mimeo.*

_ and _ , "What Motivates Effort? Evidence and Expert Forecasts," 2016. University of California, *mimeo.*

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Tool Kit," in T. Paul Schultz and John Strauss, eds., *Handbook of Development Economics, Vol. 4*, Amsterdam: Elsevier, 2008, pp. 3895–3962.

Dupas, Pascaline, "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence From a Field Experiment," *Econometrica*, 2014, *82* (1), 197–228.

Egger, Matthias, George Davey Smith, and Jonathan AC Sterne, "Uses and Abuses of Meta-analysis," *Clinical Medicine*, 2001, *1* (6), 478–484.

Ellsberg, Daniel, "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics*, 1961, *75* (4), 643–669.

Fisher, Ronald Aylmer, "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 1926, *33*, 503–513.

_ , *The Design of Experiments.*, Edinburgh and London: Oliver & Boyd, 1935.

Gilboa, Itzhak and David Schmeidler, "Maxmin Expected Utility with a Non-Unique Prior," *Journal of Mathematical Economics*, 1989, *18* (2), 141–153.

_ **, Andrew Postlewaite, and David Schmeidler**, "Is it Always Rational to Satisfy Savage's Axioms?," *Economics and Philosophy*, 2009, *25* (3), 285–296.

Gittins, John C., "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1979, *41* (2), 148–177.

Green, Donald P and Andrej Tusicisny, "Statistical Analysis of Results from Laboratory Studies in Experimental Economics: A Critique of Current Practice," 2012. Columbia University, *mimeo.*

Grossman, Sanford J and Joseph E Stiglitz, "On the Impossibility of Informationally Efficient Markets," *The American Economic Review*, June 1980, *70* (3), 393–408.

Gul, Faruk and Wolfgang Pesendorfer, "Temptation and Self-Control," *Econometrica*, 2001, *69* (6), 1403–1435.

Heckman, James J. and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, May 2005, *73* (3), 669–738.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter Van der Windt, "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration," *Political Analysis*, 2013, *21* (1), 1–20.

Jack, B Kelsey et al., "Private Information and the Allocation of Land Use Subsidies in Malawi," *American Economic Journal: Applied Economics*, 2013, *5* (3), 113–35.

Karlan, Dean, Ryan Knight, and Christopher Udry, "Hoping to Win, Expected to Lose: Theory and Lessons on Micro Enterprise Development," 2012. NBER Working Paper Series # 18325.

Karlan, Dean S. and Jonathan Zinman, "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," *Econometrica*, 2009, *77* (6), 1993–2008.

Kasy, Maximilian, "Why Experimenters Should not Randomize, and What they Should do Instead," 2013. Harvard University, *mimeo.*

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji, "A Smooth Model of Decision Making Under Ambiguity," *Econometrica*, 2005, *73* (6), 1849–1892.

Kreps, David M. and Evan L. Porteus, "Temporal Resolution of Uncertainty and Dynamic Choice Theory," *Econometrica*, 1978, *46* (1), 185–200.

Machina, Mark J, "Dynamic Consistency and Non-Expected Utility Models of Choice under Uncertainty," *Journal of Economic Literature*, 1989, *27* (4), 1622–1668.

Maynard, G.D., "Statistical Study of Anti-Typhoid Inoculation," *Biometrika*, March 1909, *6* (4), 366–375.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, January 2004, *72* (1), 159–217.

Milgrom, Paul R., "Rational Expectations, Information Acquisition, and Competitive Bidding," *Econometrica*, July 1981, *89* (4), 921–943.

Morgan, Kari Lock and Donald B Rubin, "Rerandomization to Improve Covariate Balance in Experiments," *The Annals of Statistics*, 2012, *40* (2), 1263–1282.

Muralidharan, Karthik and Vankatesh Sundararaman, "The Aggregate Effects of School Choice: Evidence from a Two-Stage Experiment," *Quarterly Journal of Economics*, 2015, *130* (3), 1011–1066.

Olken, Benjamin A, "Promises and Perils of Pre-Analysis Plans," *The Journal of Economic Perspectives*, 2015, *29* (3), 61–80.

Olszewski, Wojciech and Marcin Peski, "The Principal-agent Approach to Testing Experts," *American Economic Journal: Microeconomics*, 2011, *3* (2), 89–113.

Pearl, Judea, *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press, 2000.

Persico, Nicola, "Information Acquisition in Auctions," *Econometrica*, 2000, *68* (1), 135–148.

Robbins, Herbert, "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, September 1952, *58* (5), 527–535.

Rothschild, Michael, "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 1974, *9* (2), 185–202.

Rubin, Donald B, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 1974, *66* (5), 688–701.

Savage, Leonard J, *The Foundations of Statistics*, Courier Corporation, 1954.

Schmeidler, David, "Subjective Probability and Expected Utility without Additivity," *Econometrica*, July 1989, *57* (3), 571–587.

Simpson, R.J.S. and Karl Pearson, "Report on Certain Enteric Fever Inoculation Statistics," *The British Medical Journal*, 1904, *2* (2288), 1243–1246.

Vapnik, Vladimir, *The Nature of Statistical Learning Theory*, 2nd edition ed., Springer, 1999.