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@prinecton.edu
princeton.edu/∼chassang/

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

April 4, 2015

Abstract

We discuss possible decision theoretic foundations for experiment design and argue that a non-Bayesian minimax framework is best suited to capture the objectives of actual experimenters. We use this framework to explore questions of optimal design, and external validity. We show that randomization and frequentist decision-making obtain naturally in internal decision making problems. However, external decision making remains inherently subjective. We embrace this conclusion and propose a framework for structured speculation that: (i) provides experimenters a formal way to express their beliefs about the external validity of their findings; (ii) exploits subjective beliefs in a robust way.

JEL Classifications: C93, D70, D80

Keywords: experiment design, non-Bayesian experimentation, randomization, self-selection, external validity, ambiguity aversion

*Prepared for the *Handbook of Field Experiments*. Chassang and Snowberg gratefully acknowledge the support of NSF grant SES-1156154.

1 Introduction

We seek to provide a decision theoretic framework in which to study the problem of optimal experiment design. Clearly, experimentation and learning have been widely studied by economic theorists (Grossman and Stiglitz, 1980; Milgrom, 1981; Banerjee, 1992; Persico, 2000; Bergemann and Välimäki, 2002). Bandit problems, in particular, are used to formalize the problem of optimal experimentation, and specifically capture the tradeoffs between experimentation and exploitation (Robbins, 1952; Bellman, 1956; Rothschild, 1974; Gittins, 1979; Aghion et al., 1991; Bergemann and Välimäki, 1996, 2006)

Yet there seems to be almost no connection between this work and the empirical literature based on experimental or quasi-experimental approaches. The primary concern in this literature, as far experimental design is concerned, has been with making sure that the treatment and the control samples are balanced in the sense of not being systematically different. Conditional on their being balanced, the next concern is with power to discrminate between alternative hypotheses.

From a purely Bayesian perspective, this prioritization is hard to justify. A Bayesian is able to learn from both treatment and control even if they are very different, because she has the additional benchmark of her prior and therefore may very well be willing to trade-off balance for additional power. As a result, as we will highlight in Section 3 of this chapter, a Bayesian typically will strictly prefer to assign treatments in a deterministic way. The practice, on the other hand among experimenters is to assign them randomly, for the intuitive reason that this ensures robust balance.

The rest of section 3 is therefore devoted to reconciling these two perspectives. We point out that there are good reasons why a policy experimenter may not want to embrace the pure Bayesian stance, and suggest that a good first step towards modeling an experimenter is to assume that she is a maximin decision-maker in the tradition of (among others) Schmeidler (1989); Gilboa and Schmeidler (1989); Klibanoff et al. (2005).

Following Banerjee et al. (2015), we show that in the maximin decision framework under one important but plausible assumption, randomizing is always optimal as long as the sample size is large enough and the decision-maker is sufficiently ambiguity averse. Balance is important because it limits the possible set of alternative interpretations of the evidence. We suggest that this is the view that the experimental literature has implcitly taken.

Section 4 explores the implications of the maximin decision framework for the design of experiments. We follow Banerjee et al. (2015) in showing that in the maximin framework, unlike in a Bayesian world, pre-registration of experimental designs may be optimal. On the other hand, it also tells us that the experimental practice of trying to eliminate all forms of selection is often sub-optimal–experimenters are in effect throwing away information. By following Chassang et al. (2012), we show that selection and randomization can be accommodated in the same experimental design.

The more speculative final sections are dedicated to the question of how to improve the external validity of experiments. We show that external decision making is necessarily subjective. The penultimate section embraces this conclusion and develops a framework for structured, informed, speculation. Practically we suggests that experimental papers include a section allowing authors to speculate (in a rigorous way) about how the treatment they study may work in other environments. We emphasize the potential role of experimental registries in archiving, encouraging, and disciplining such speculation. We also suggest ways in which policymakers can robustly exploit such subjective assessments.

The concluding section discusses, qualitatively, how a system for structured speculation might work in practice. In particular, we discuss common dimensions of external validity to be addressed: namely, how a treatment would scale, and how it would affect groups of particular interest to researchers. We further discuss how this applies to questions of interest to development economists: gender, inter-ethnic relationships, participant beliefs, and credit constraints. Moreover, we discuss how an increased focus on experimental validity may push experimenters and theorists to develop and use tools for better experimental design.

The rest of this section very briefly discusses the history of thinking about experimental design.

History

The first documented controlled experiment is found in the biblical book of Daniel, a story set around 605 B.C.E., to compare 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)

Yet, despite the early and intuitive appeal of controlled experiments, it took statistical theorists experienced with field experiments to insert randomization into the process. 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 that position, citing 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).

A large statistics literature on experiment design followed this lead, and is summarized in **?**. The primary question of interest is how to assign a limited sample of observations across

different covariates to minimize some version of prediction error. A particular version of this objective, called G-optimality, involves minimizing the maximum variance of the predicted values across the different values of the covariates. This is related in spirit to the maximin approach we adopt here.

While much of this work has foundations in classical statistics, there is also work in statistics on Bayesian experimental design. 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 Setup

Throughout this chapter we consider the problem of choosing whether or not to implement a policy (or treatment) given the information from an experiment. Formally, the person who makes the decision over policy implementations will be the same person who designs the experiment to inform that decision. At various points in the chapter it may be useful to think of an experimenter whose policies recommendations need to be justified to an audience or stakeholders.

Our notation follows that of Banerjee et al. (2015). We consider a decision-maker that needs to decide on whether to implement some policy $a \in \{0, 1\}$, that provides a given treatment $\tau \in \{0, 1\}$ to a unit mass of individuals indexed by $i \in [0, 1]$.[1] To inform her judgement, the decision-maker is able to run experiments assigning a given number N of subjects to treatment and control.

Potential outcomes for subject i as a function of treatment τ are denoted by $Y_i^\tau \in \{0, 1\}$. Event $Y = 1$ is referred to as a success. Each individual i is associated with covariates $x_i \in X$, where 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.

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

Outcomes Y_i are i.i.d. conditional on covariates. We denote by $p_x^\tau \equiv \text{prob}(Y_i^\tau = 1 | x_i = x)$ the probability that $Y_i^\tau = 1$ conditional on covariate x.

Environments and Decision Problems. An environment z is described by the finite-dimensional vector p of success probabilities conditional on covariates,

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

For much of what follows we consider only a single environment p, and suppress the subscript z when it is not important and will not cause confusion (i.e. until Section 5).

Given an environment p and a policy decision $a \in \{0,1\}$, the decision-maker's payoff $u(p,a)$ can be written as

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

In general, the *experimental environment*, denoted by z_0, need not be the same as a *policy environment*, denoted by $z_k,\ k \geq 1$. If a policy environment is the same as the experimental environment, $p_{z_k} = p_{z_0}$, then we say that the decision maker faces an *internal* decision problem. Otherwise, the policy maker faces an *external* decision problem. This allows us to address external validity in Section 5.[2]

Experiments and Decision Rules. An experiment is a tuple $e = (x_i, \tau_i)_{i \in \{1,\ldots,N\}} \in (X \times \{0,1\})^N \equiv E$. Experiments generate outcome data $y = (y_i)_{i \in \{1,\ldots,N\}} \in \{0,1\}^N \equiv \mathcal{Y}$, with y_is independent realizations of $Y_i^{\tau_i}$ given (x_i, τ_i).

The decision-maker's strategy consists of both an experimental design $\mathcal{E} \in \Delta(E)$ and an decision rule $\alpha : E \times \mathcal{Y} \to \Delta(\{0,1\})$ mapping observable experimental outcomes (including the realized experiment design e) to policy decisions a. We denote by $\mathcal{A}$ the set of possible

[2]Note that the environment may differ because either the population has different baseline outcomes, and/or different responses to treatment, or because the population is the same but the treatment differs in some way (or both).

decision rules.

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).

Definition 1 (equivalent experiments)**.** *We say that two experiments* $e = (x_i, \tau_i)_{i\in\{1,\dots,N\}}$ *and* $e' = (x'_i, \tau'_i)_{i\in\{1,\dots,N\}}$ *are* equivalent*, denoted by* $e \sim e'$*, if there exists a permutation* $\sigma : \{1,\dots,N\} \to \{1,\dots,N\}$ *of the subjects' labels such that* $(x_i, \tau_i) = (x'_{\sigma(i)}, \tau'_{\sigma(i)})$ *for all* i*. The equivalence class of an experiment* e *is denoted by* $[e]$*.*[3] *We denote by* $[E]$ *the partition of experiments* $e \in E$ *in equivalence classes. We say that two experimental designs* $\mathcal{E}$ *and* $\mathcal{E}'$ *are equivalent, denoted* $\mathcal{E} \sim \mathcal{E}'$*, if they induce the same distribution over* $[E]$*.*

Discussion. The framework we propose here does not aspire to any great generality. The goal is to provide us with just enough flexibility to illustrate the issues that interest us. We do not consider, for example, more sophisticated decision rules which allow different actions to be chosen for different subsets of the covariates based on the same experiment. We also assume that treatment and control observations are substitutable and the only constraint is on the total number of observations. Both these assumptions are easily relaxed.

3 Frameworks for Experimental Design

Many frameworks have been proposed to think about the problem of experimental design. In this section we summarize the two main decision theoretic approaches

3.1 Bayesian Experimentation

Much of economic theory proceeds under the assumption that decision makers are subjective expected utility maximizers. Formally, let the decision-maker start from a prior $h_0 \in \Delta(\mathcal{P})$,

[3] It is convenient to include distributions $\mathcal{E}$ with support in $[e]$ in the equivalence class of e.

In the context of our optimal 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 (Bayesians do not randomize)**.** *Assume that the decision-maker is Bayesian, that is, designs experiments according to* (1)*.*

1. *There exist deterministic solutions $e \in E$ to* (1)*. A mixed strategy $\mathcal{E} \in \Delta(E)$ solves* (1) *if and only if for all $e \in \text{supp } \mathcal{E}$, e solves* (1)*.*

2. *For generically every prior h, there exists a unique equivalence class of strictly optimal experiments $[e]$.*

The intuition for this result comes from the fact that a Bayesian decision-maker wants to assign each observation to maximize the impact on her posterior beliefs given her current prior. For generic priors there is a unique optimal choice for each observation. The formal proofs of these results are given in Banerjee et al. (2015).

Kasy (2013) argues from Proposition 1 (i) that randomized controlled trials are suboptimal. The intuition he provides for the result is different and more akin to arguments from classic statistics. His point is that if the goal is to achieve balance between the treatment and control samples, this is more efficiently done by purposefully assigning particular observations 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.

It is also worth commenting that the result does not say that all observations with the same covariate should be assigned either to treatment or control. It may well be that some of the observations at a particular x may be treatment and others control, and in

this case, there are multiple multiple equivalent experiments and the proposition does not rule out randomizing between them, in effect assigning different observations with the same covariates to treatment and control by lottery is not ruled out, but has no advantage either.

To us these results say that Bayesian expected utility maximization is obviously limited as a suitable paradigm for understanding what experimenters are after. More starkly, if Bayesians do not randomize, then the choice-theoretic foundations of modern experimentation must not be Bayesian. But what, then, is an appropriate decision-theoretic model for experimentation?

3.2 Ambiguity, or an Audience

Subjective expected utility maximization has strong normative foundations and the "as if" axiomatization proposed by Savage (1954) seem so natural that subjective expected utility maximization is sometimes considered to be an expression of rationality. However, there are several reasons to question the applicability of this paradigm to questions of experimental design.

First, our 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 completness axiom (Gilboa et al., 2009; Bewley, 1998). Third she may recognize that she needs to convince others whose priors may diverge from her own. Finally, she may worry about being accused of fraudent manipulation of the evidence.

One possible reaction to the need to "defend against" alternative priors is to require Blackwell Dominance. Blackwell (1951) defines an ordering of data, and thus experiments, that is robust to a decision-maker's prior and preferences. An experiment $\mathcal{E}$ Blackwell dominates an experiment $\mathcal{E}'$ if it allows any subjective expected utility maximizer to achieve a higher expected payoff, regardless of their preferences and prior belief over the policy

environment. Whenever an experiment Blackwell-dominates another, it is preferred by all Bayesian decision makers.

Definition 2 (Blackwell Dominance)**.** *An experiment $\mathcal{E}$ is said to Blackwell dominate an experiment $\mathcal{E}'$ if and only if for all priors $h \in \Delta(P)$ over states of the world p, and all utility functions $v(a,p)$,*[4]

$$\max_{\alpha \in \mathcal{A}} \mathbb{E}_{h,\mathcal{E}} \left[v(\alpha, p) \right] \geq \max_{\alpha \in \mathcal{A}} \mathbb{E}_{h,\mathcal{E}'} \left[v(\alpha, p) \right].$$

Blackwell dominance is a very strong requirement, and Blackwell's theorem establishes that it can be obtained in only one way: an experimental design $\mathcal{E}$ Blackwell dominates an experimental design $\mathcal{E}'$ if and only if the joint distribution over states of the world and data induced by $\mathcal{E}'$ can be obtained by garbling the observable outcomes produced by experiment $\mathcal{E}$.

Theorem 1 (Blackwell (1951))**.** *Experiment $\mathcal{E}$ Blackwell dominates experiment $\mathcal{E}'$ if and only if there exists a garbling, i.e. a mapping $g : E \times \mathcal{Y} \to \Delta(E \times \mathcal{Y})$, such that for all states of the world p, the distributions of observable (e, y) under $\mathcal{E}$ and $\mathcal{E}'$ satisfy*

$$(e, y)|p, \mathcal{E}' \sim g(e, y)|p, \mathcal{E}.$$

The main drawback of Blackwell's formulation is that it produces only a partial order on experiments. That is, given two experiments that are Blackwell-undominated by each other, a different decision maker with a different prior (or set of priors) will often rank those experiments differently in terms of their perceived efficiency for decision making. Therefore, while we will make use of Blackwell-dominance wherever possible (for example in Section 4.3), it is not suitable as a basis for a general approach to experimentation.

Instead we follow Banerjee et al. (2015) and propose a model of ambiguity averse experimentation. Under mild assumptions, the maximin expected utility framework of Gilboa

[4] Although all of this chapter deals with the specific utility $u(a,p)$ defined earlier, we use v here to indicate that decision makers may have different preferences over outcomes.

and Schmeidler (1989) can be expressed as follows. The decision-maker chooses experiment design $\mathcal{E}$ and allocation rule α solving

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

where $\lambda \in [0,1]$. As before, h_0 is a fixed prior and H is a convex set of priors $h \in \Delta(P)$. The fact that the decision-maker seeks to maximize her utility given a prior that would tend to minimize it lends this framework the title of "maximin utility".

While one could take the inspiration of this model literally and assume the decision-maker is ambiguity averse, Banerjee et al. (2015) emphasize an alternative interpretation, which may be particularly compelling when the purpose of experimentation is to influence policy. The idea is that h_0 represents the fixed prior of the decision maker, and H represents the set of possible priors in an audience. The parameter λ can then be seen as the the weight that the decision-maker places on her preferences versus those of the audience. Note that if $\lambda = 1$, we recover (1), so 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.

This formulation does not directly deal with the possibility, mentioned above, that Savage's completeness postulate may fail. However, as we will point out below, there is a sense in which it does address this criticism. Finally, while we do not formally deal with the possibility of "moral hazard" on the part of the decision-maker, we discuss it briefly in the concluding section.

Note that although maximin utility allows for a complete ordering of experiments. A given decision-maker can always rank two possible experimental designs in terms of their efficiency for decision making. However, that ranking will often depend on the prior and preferences of the particular decision maker, and a different decision-maker with a different prior (or set of priors) and/or preferences may have a different ordering.

4 Ambiguity Averse Experimentation

We now develop the implications of our framework for optimal experimentation. We emphasize features of experiment design for which there are discrepancies between practice and standard theory.

4.1 Why randomize?

Following Banerjee et al. (2015), we argue that the framework described in Section 3.2 is not only compatible with experimenters randomizing, but explains several stylized facts on when experiment choose to randomize, or experiment deliberately (i.e. deterministically).

Denote by $\overline{p}^a \equiv \sum_{x \in X} q(x) p_x^a$ the expected probability of success, given policy $a \in \{0,1\}$. For $X_0 \subset X$, we denote by $p_{X_0} \equiv (p_x^\tau)_{\substack{x \in X_0, \\ \tau \in \{0,1\}}}$ the subset of success rates for covariates $x \in X_0$. Then, the following *limited extrapolation* condition implies that experimenters randomize whenever the sample size is large enough.

Assumption 1 (Limited Extrapolation I)**.** *We assume that there exists $\nu > 0$ such that, for all $X_0 \subset X$ with $|X_0| \leq N$, 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},*

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

This implies that knowing success rates at any subset of N covariates is not a sufficient statistic for efficient internal decision making: the residual uncertainty over which policy maximizes population-level outcomes remains bounded away from 0.[5] In particular, even if the audience were to learn the state at N values of the covariate x, some member of the audience would still maintain sufficient belief that the treatment could have a substantial positive or negative average treatment effect (greater than ν).

[5]Note that Assumption 1 implies that $N < |X|$.

With this assumption imposed, we report the following result from Banerjee et al. (2015).

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 random. In particular, for any N, the optimal experiment $\mathcal{E}^*$ satisfies

$$\max_{\alpha} \min_{h \in H} \mathbb{E}_{h,\mathcal{E}^*} \left[u(\alpha(e,y),p) \right] \geq \min_{h \in H} \mathbb{E}_h \left(\max_{a \in \{0,1\}} p^a \right) - \frac{\sqrt{N} \exp(-1/2)}{N-1},$$

whereas, for any N, all deterministic experiments $e \in E$ are such that

$$\max_{\alpha} \min_{h \in H} \mathbb{E}_{h,e} \left[u(\alpha(e,y),p) \right] < \min_{h \in H} \mathbb{E}_h \left(\max_{a \in \{0,1\}} p^a \right) - \nu.$$

Proposition 2 suggests that optimal experimentation depends crucially on how much weight the decision-maker places on satisfying an audience, and on the availability of sample points. Point (i) shows that when sample points are expensive, or when the decision-maker does not put much weight on satisfying an audience (λ close to 1), optimal experimentation will be Bayesian. That is, the experimenter will focus on sampling the subjects from whom she learns the most. Point (*ii*) shows that when sample points are cheap and the decision-maker cares about satisfying a sufficiently adversarila audience, she will use randomized trials which allow for a prior-free identification of correct policies.

This proposition broadly agrees with stylized facts about experimentation: when the experimenter is concerned with satisfying a skeptical audience, as in much of scientific researc, she randomizes. However, a firm, when trying a new production process, will not randomize

which plants try it out. It will focus on a few plants where it feels it can learn the most about the process and its benefits. But even firms randomize in their on-line marketing campaigns, where data points are extremely cheap.

One implication of Proposition 2 (ii) worth emphasizing, is that a decision-maker who adopts the randomization protocol without undersatnding all its ramifications (why it is being chosen, what audience it is meant to satisfy) will nevertheless do almost as well as she ever could as long as N is large enough. Specifically even if someone (or her own doubting mind) comes up with a particularly unfriendly prior, she will still be satisfied with the decision rule adopted. In this sense this addresses the concern that decision makers may violate Savage's completeness postulate.[6]

To get some intuition about the result, consider a setting where the decision-maker's prior treats all the covariates as being exactly identical. Therefore she would be happy to assign treatment to any subset of X. However, she is concerned that a skeptical audience would take the view that those particular $x's$ are (say) particularly favorable to treatment and since this could happen for any subset of X, she would be better off randomizing and therefore avoid being accused of favoring any specific $x's$.

The one caveat to this explanation is that she may be able avoid the same accusation by choosing the right set of x's especially if there is enough continuity in p as a function of x(say if you assign treatment to any x, you also control to some other x'very close to it).[7]This is what our assumption of Limited Extrapolation I rules out. Essentially the idea is that there is someone in the potential audience who skeptical enough to question any continuity with respect to x, which seems likely.

Finally Kasy (2013) reports a result that on the surface directly contradicts this proposition– the claim is that randomized experiments can never do better than deterministic ones even

[6]A similar results hold for more complex policies that vary treatment with covariate X, provided that the complexity of possible policies is limited. See Vapnik (2000) for a reference.

[7]This is related to the point made by Kasy (2013) about the possibility of highly efficient deterministic experiments.

with a maximin objective. This is however an artifact of the assumption that the audience's priors react to the realization of the randomization: i.e. the audience gets to see the realized sample before picking its prior. One way to justify this kind of extreme skepticism would be to assume that the randomization is not public and that the decision-maker re-randomizes till she gets the desired distribution of covariates in treatment and control. The appeal of randomized trials suggests that at least from the point of view of audiences, this danger is exaggerated. However it does make a case for making the randomization process as transparent as possible. This relates to the question of experimental registration.

4.2 Why Register Experiments?

As Banerjee et al. (2015) observe, the value that practicing experiment designers place on registering experiments is another challenge for existing theory. It is shown that while registration is never optimal for a Bayesian decision-maker, it may be optimal for an ambiguity averse decision-maker with preferences described by (2).

Formally, unregistered experiments are captured as dynamic experimentation problems without commitment. Experimental subjects are chosen sequentially, after observing past data: for each $t \in \{1, \cdots, N\}$, the selection rule $\mathcal{E}_t$ for the t^{th} subject's characteristics and treatment status $e_t = (x_t, \tau_t)$ is measureable with respect to past data $(e_s, y_s)_{s \in \{1, \cdots, t-1\}}$.

Proposition 3 (Banerjee et al. (2015))**.** *Assume that $\lambda = 1$, so that the decision-maker is Bayesian. Then:*

(i) Generically over priors h_0, it is strictly optimal for a Bayesian decision-maker not to register her experiment.

(ii) Posterior belief $h_0(p|e, y)$ does not depend on the experiment $\mathcal{E}$ which was used to obtain data (e, y).

The first part of the proposition follows from the fact that a Bayesian would use data from past experiments (or the current experiment up to time t) in order to optimally select

the covariates and treatment statuses of subjects in the next phase of their experiment. This follows from Proposition 1. As registering an experiment would limit the Bayesian's ability to optimally sample at each point in time, they would strictly prefer to not register. The second part of the proposition implies that a Bayesian does not learn differently from the results of registered and unregistered experiments. Provided all the data is made available, data is treated the same no matter how it is generated, and thus registration gives no more information than ex post examination of the data.

In contrast, ambiguity-averse decision makers, or a decision-maker facing an audience, may wish to register an experiment, provided that sufficiently many data points are available. This comes from the fact that ambiguity-averse decision makers are not time consistent.

Consider the problem of a forward-looking decision-maker. Denote histories or realized experiments by $\phi_t = (e_s, y_s)_{s<t}$ and expectations about future behavior $\mathcal{E}^t = (\mathcal{E}_s)_{s>t}$. Without registration, the decision maker's policy choices $(\mathcal{E}_t^*, \alpha^*)_{t\in\{1,\cdots,N\}}$ are such that

$$\mathcal{E}_t^* \in \underset{\mathcal{E}_t\in\Delta(X\times\{0,1\})}{\operatorname{argmax}} \lambda\mathbb{E}_{h_0}\left(u(\alpha^*,p)|\phi_t,\mathcal{E}_t,\mathcal{E}^{*,t},\alpha^*\right) + (1-\lambda)\min_{h\in H}\mathbb{E}_h\left(u(\alpha^*,p)|\phi_t,\mathcal{E}_t,\mathcal{E}^{*,t},\alpha^*\right)$$

$$\alpha^*(e,y) \in \underset{a\in\{0,1\}}{\operatorname{argmax}} \min_{h\in\widehat{H}} \mathbb{E}_h\left(u(a,p)|z_N\right).$$

In contrast, if the experiment is registered, the decision-maker's problem coincides with the original one studied in (2), i.e. $(\mathcal{E}^*, \alpha^*)$ solve

$$\max_{\mathcal{E},\alpha} \lambda\mathbb{E}_{h_0,\mathcal{E}}[u(\alpha(e,y),p)] + (1-\lambda)\min_{h\in H}\mathbb{E}_{h,\mathcal{E}}[u(\alpha(e,y),p)].$$

Proposition 4 (Banerjee et al. (2015))**.** *Assume that* $\lambda < 1$.

(i) Take N as given. For generically every h_0, there exists $\epsilon > 0$ such that for $\lambda > 1-\epsilon$, it is optimal not to register.

(ii) Take λ and h_0 as given. There exists $\underline{N}$ such that whenever $N \geq \underline{N}$, it is

optimal for the experimenter to register.

The first part of the proposition can be seen through "continuity" with the first part of Proposition 3—in the limit as $\lambda \to 1$, the two are the same. The more interesting result is the second, which is the result of the time-inconsistency noted above. In particular, the prior that will give the minimum payoff to the decision-maker at a given point in time will be different from the minimizing prior when all the data is collected. The decision maker will use this locally-minimizing prior to pick her next sample. As a result, she will continually optimize using priors that result in poor choices from the point of view of the globally-minimizing prior. If the decision-maker could commit to a sample ahead of time, this would thus raise the payoff of the decision-maker.

Further work. The adversarial model of Banerjee et al. (2015) provides a rationale for why it may be optimal to commit ex ante to a sampling frame: this minimizes the impact of disagreement on information acquisition and decision making. However, even a decision-maker facing a skeptical audience would not find any use in filing a pre-analysis plan, nor would her audience care if she did or did not as long as all the data is made available. This occurs as the decision-maker—although non-Bayesian—and her audience can still process all available data at no cost. We believe that a model of costly information processing à la Sims (2003) would provide foundations for the value of pre-registering hypotheses. A careful formalization of such a model is beyond the scope of this paper, but we believe it is a useful objective for future research, since it may help guide the design of experiment registration systems. Another useful application for our framework would be to assess the consequences of rerandomization.

4.3 Is Self-selection Bad?

Self-selection is typically thought of as a design flaw in trial design, but several recent studies have made the point that charging subjects for treatment lets researchers elicit treatment ef-

fects conditional on unobserved values (Karlan and Zinman, 2009; Ashraf et al., 2010; Cohen and Dupas, 2010). Chassang et al. (2012) formalize this point and show that correctly designed selective trials can Blackwell-improve standard randomized controlled trials. They are interested in environments where the subjects' behavior may affect treatment outcomes, and behavior may be different across environments. In such environments, since effort depends on beliefs, treatment effects in the same population may well vary over time.

A simple example. Assume that an infinite sample $N = +\infty$ is available. Chassang et al. (2012) illustrate their point with a stylized model, but their results hold quite generally. Imagine that treatment corresponds to providing the subject with a new technology, and that the probability of a successful outcome $Y = 1$ depends on both treatment status $\tau \in \{0, 1\}$ and on an effort decision $e \in [0, 1]$

$$\text{Prob}(y_i = 1 | e_i, \tau_i) = \rho + R e_i \tau_i,$$

where $e_i \in [0, 1]$ is subject i's decision of whether or not to expend effort using the technology. Parameter $R \in [R_L, R_H]$ is the technology's return—which is common to all subjects—and ρ is the unknown baseline likelihood of success over the study period. For simplicity, we ignore observable characteristics $x \in X$, i.e. presume that subjects are homogeneous on observable characteristics. Instead, we let subjects have different beliefs about returns R. We denote by R_i^e the returns expected by a subject i. The distribution F_{R^e} of expectations R^e in the population need not be known to the principal or the subjects.

Given effort e_i, subject i's expected utility is given by

$$\mathbb{E}_i[y_i | e_i] - c e_i,$$

where $c \in (R_L, R_H)$ is the subjects' cost of effort. In addition, we assume each subject has quasilinear preferences with respect to money. A subject's willingness to pay for treatment

is $V_i = \max\{R_i^e - c, 0\}$, which we assume is less than some value $V_{\max}$ for all subjects.

We focus initially on open trials, in which subjects know their treatment status, and contrast two ways of running trials: a standard RCT, where subjects are randomly assigned to the treatment group with probability q, and a selective open trial that allows subjects to express preferences for treatment by selecting their probability of treatment.

A focal implementation of selective trials uses the BDM mechanism (see Berry et al., 2012, for a detailed field implementation), and proceeds as follows:

1. Each subject sends a message $m_i \in [0, V_{\max}]$ indicating her willingness to pay for treatment;

2. A price π_i to obtain treatment is independently drawn for each subject from a distribution with convex support and c.d.f. F_π that satisfies $0 < F_\pi(0) < F_\pi(V_{\max}) < 1$; and

3. If $m_i \geq \pi_i$, the subject obtains the treatment at price π_i, otherwise, the subject is in the control group and no transfers are made.

Note that a higher message m increases a subject's probability of treatment, $F_\pi(m)$, as well as her expected payment: $\int_{\pi \leq m} \pi \, \mathrm{d}F_\pi$. Since F_π has convex support, it is dominant for a subject of type t to send message $m = V_i$. Selective trials essentially let subjects pick their likelihood of being treated at some cost. Standard randomized controlled trials are trials in which the likelihood of being treated is fixed.

Proposition 5 (Chassang et al. (2012))**.** *Selective trials Blackwell-dominate randomized controlled trials.*

The reason why this holds is that selective trials make self-selection explicit, while still constructing a treatment and sample pool conditional on preferences.

Inference is particularly straightforward in the context of this simple example. Consider first the standard RCT. If subject i is in the treatment group, he chooses to expend effort

$e = 1$ if and only if $R_i^e \geq c$. Hence, the average treatment effect identified by an RCT is

$$\begin{aligned}
\Delta^{RCT} &= \mathbb{E}[y|\tau = 1] - \mathbb{E}[y|\tau = 0] \\
&= \mathbb{E}[\rho + R \times \mathbf{1}_{R_i^e \geq c}|\tau = 1] - \mathbb{E}[\rho|\tau = 0] \\
&= R \times \text{Prob}\left(R_i^e > c\right) = R \times (1 - F_{R^e}(c)).
\end{aligned}$$

When the distribution of subjects' expectations F_{R^e} is known, then an RCT will identify R. However, in most cases F_{R^e} is not known, and the average treatment effect Δ^{RCT} provides a garbled signal of the underlying returns R. If the outcomes in the treatment group are only weakly better than those in the control group, the principal does not know if this is because the new technology is not particularly useful, or because the subjects did not expend sufficient effort using it.

Selective trials elicit the subjects' willingness to pay, and, conditional on a given willingness to pay V, generates non-empty treatment and control groups. Since it is dominant for subjects to truthfully reveal their value, a subject with value V_i has probability $F_\pi(V_i)$ of being in the treatment group and probability $1 - F_\pi(V_i)$ of being in the control group. Both of these quantities are strictly positive as $0 < F_\pi(0) < F_\pi(V_{\max}) < 1$. This lets us estimate marginal treatment effects (MTEs, Heckman and Vytlacil, 2005). That is, for any willingness to pay V, we are able to estimate

$$\begin{aligned}
\Delta^{MTE}(V) &\equiv \mathbb{E}[y|\tau = 1, V_i = V] - \mathbb{E}[y|\tau = 0, V_i = V] \\
&= \mathbb{E}[y|\tau = 1, m_t = V] - \mathbb{E}[y|\tau = 0, m_t = V],
\end{aligned}$$

which can be used to perform policy simulations in which the distribution of types is constant but access to the technology is changed, for example, by subsidies. Moreover, MTEs can be integrated to recover the average treatment effect identified by an RCT.

In the current environment, willingness to pay is also a good signal of future use, and thus

MTEs can be used to identify the true returns R. Specifically, all subjects with value $V_i > 0$ have expectations R_i^e such that $R_i^e - c > 0$, and expend effort $e = 1$ using the technology. Hence, it follows that

$$\begin{aligned} \Delta^{MTE}(V > 0) &= \mathbb{E}\left[\rho + R \times e_t \,|\tau = 1, V_i > 0\right] - \mathbb{E}[\rho|\tau = 0, V_i > 0] \\ &= R. \end{aligned}$$

A selective trial identifies the average treatment effect, MTEs, and true returns R. Hence, it is more informative than an RCT, which only identifies the average treatment effect.

The true returns R, and the distribution of valuations V_i, are useful in several external decision-making problems. First, knowing R allows us to simulate the treatment effect for a population in which all subjects expend the appropriate amount of effort. Second, these variables allow us to estimate the returns to increasing usage within a given population. Third, the data provided by a selective trial can be used to inform subjects and disrupt learning traps more effectively than data from an RCT. For example, imagine that true returns to the technology are high, but most subjects believe they are low. In that case, an RCT will measure low returns to the treatment and will not convince subjects that they should be expending more effort. In contrast, the data generated by a selective trial would identify that true returns are high, and lead subjects to efficiently adopt the new technology. Finally being able to estimate the treatment effect at different levels of valuation, allows us perform optimal subsidy computation (Heckman and Vytlacil, 2005).

Limitations of selective trials. The result that selective trials Blackwell-dominate randomized controlled trials is very strong: it applies for any given subjective belief over the underlying environment, and does not rely on the specific illustrative model described above. However, these results rely on the fact that an arbitrarily large sample is available.

For small samples, running selective trials may come at the cost of power for the follow-

ing reason: any allocation mechanism that elicits subjects' valuations in strictly dominant strategies must treat subjects with higher valuations with strictly higher probability. This may come at the cost of power. This has practical consequences. Consider for instance the BDM mechanism described above. If the true distribution of values is concentrated at the bottom of the value interval $[0, V_{\max}]$, the BDM mechanism will result in a large fraction of the sample being placed in the control group (see Cohen and Dupas, 2010; Dupas, 2014, for illustrations of sampling costs induced by self-selection). The monotonicity requirement described above means that some level of deviation from the treatment/control division of observations that maximizes power is unavoidable.

An implication is that successful implementation of selective trials in small sample environments requires careful piloting. In particular, estimating the demand curve for treatment is needed to design trials with appropriately balanced samples.

Another potential limitation of selective trials is that, as in our example above, we need a model to figure out exactly what to take away from these results. If the "production function," i.e. the mapping between valuation, effort and returns was unknown, we would clearly learn less from the experiment. However the result about Blackwell dominance still holds in the sense that we know strictly more about the distribution of returns as a function of the valuation; this can be useful for example if we have a potential intervention that moves the distribution of valuations in a way that we can measure–a subsidy for example–even if we known nothing about the "production function".

5 External Validity

Section 4.1 showed that randomization is the solution to internal decision problems when the decision-maker's audience becomes sufficiently adversarial. We now argue that decision rules for external environments may be unavoidably Bayesian, especially as the audience becomes adversarial. However, eventhough external recommendations are necessarily subjective, we

show that, by using proper incentive schemes, they can at least be made to be honest best guesses. Moreover, provided that there are several opportunities to implement the same or related treatments, subjective recommendations can be safely exploited without assuming that experimenters' subjective posteriors are correct. This section proposes a formalized way for experimenters to express their beliefs about policy outcomes in external environments, and specifies the technical details that could be useful in such a system.

5.1 External Policy Advice is Bayesian

Recall that Section 2 established the idea of different environments $z \in Z$, characterized by state p_z. As in that section, z_0 is the experimental environment, and z_k is any other environment. In general, we assume $p_{z_k} \neq p_{z_0}$ so that the experimenter has to make recommendations for an external problem. To discuss inference and recommendations in external environments, we define $H_{|p_z}$ as the set of marginal distributions $h_{|p_z}$ for $h \in H$. A key input is the degree to which information about environment z_0 reduces the amount of ambiguity aversion over other environments. We assume that the audience is sufficiently adversarial in the following sense. to cover external environments.

Assumption 2 (Limited Extrapolation II)**.** $H_{|p_{z_0}} \times H_{|p_{z_k}} \subset H$.

This implies that information obtained in environment z_0 does not restrict (although it may change) the set of beliefs that can be entertained by the audience in environment z_k. This is clearly a somewhat extreme separation assumption, but it captures cleanly our point that between what greater ambiguity aversion has very different consequences in internal and external validity problems. In fact a bounded rationality argument suggest that Assumption 2 may be plausible even if it is possible to learn about environment z_1 using data from environment z_0. Indeed, it is very difficult at the time of experiment design to anticipate which data that needs to be collected in order to address the future concerns over external validity that audience members may have. In a sense, external environments are the

environments that are not (or cannot) be taken into account when designing the experiment.

After running an experiment in environment z_0, the experimenter is asked makes a recommendation for the external environment z_k. Formally, the experimenter chooses $\mathcal{E}$ and α solving

$$\max_{\mathcal{E},\alpha} \left\{ \lambda \mathbb{E}_{h_0}[u(\alpha, p_{z_k})] + (1-\lambda) \min_{h \in H} \mathbb{E}_h[u(\alpha, p_{z_k})] \right\}. \tag{3}$$

Proposition 6. *The optimal recommendation rule α^* in (3) depends only on the experimenter's posterior belief $h_0(p_{z_k}|e)$ given experimental realization e. The optimal experiment $\mathcal{E}^*$ is Bayesian optimal under prior h_0.*

This implies that 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_0 does not change the set of priors entertained by the audience, that is, it does not reduce the ambiguity in environment z_k. Hence, differences in policies driven by experimental outcomes only reflect changes in the experimenter's own subjective beliefs. This further implies that the most information one can hope to obtain (in a Blackwell sense), is the experimenter's posterior belief over state p_{z_k}.

5.2 Structured Speculation

Proposition 6 implies that external policy advice is an unavoidably subjective enterprise. However, this does not imply that subjective recommendations by experimenters must be undisciplined, or that they should be blindly trusted. Drawing on the literature on incentivizing experts (see, for example, Olszewski and Peski, 2011; Chassang, 2013), we propose a framework to: (i) elicit truthful reports from experimenters; (ii) use these reports in a robust way that does not leave policy makers at the whims of experimenters' subjective beliefs.

Practically, we would like to encourage experimental papers to include a "Speculation" section in which authors report the mean of their posteriors in environments that they deem relevant. Practically speaking, this means a list of environments where the authors believe

that the treatment they study would, or would not, "work". We provide a discussion of useful environments for authors to address in Section 6. Experimental registries could be leveraged for this purpose by allowing authors to register, and modify their posterior beliefs regarding treatment effects in external environments.

Eliciting posteriors: To state this proposal formally, we extend the definition of $\overline{p}^a$, the average effect of policy a to cover all external environments z_k: $\overline{p}^a_z \equiv \sum_{x \in X} p^a_{x,z} q_z(x) \in [0,1]$. A experimenter's belief $h(\overline{p}^a_z|e)$ over treatment effects is a high-dimensional object. Thus, as an initial step, we propose to elicit its first moment and discuss the possibility of other moments below. Denote by $\mu_k(a) \equiv \mathbb{E}_{h_0}[\overline{p}^a_{z_k}|e]$ the mean of the expert's posterior over $\overline{p}^a_z$.

Take as given K implementations of policies $a_1, \cdots, a_K$ in environments $z_1, \cdots, z_K$ yielding data $y_1, \cdots, y_K$, with sample sizes $N_1, \cdots, N_K$, and iterative predictions $\mu_1, \cdots, \mu_K$ from the decision-maker. Define reliability score Ψ as follows

$$\forall k \in \{1, \cdots, K\}, \quad \mathcal{L}_k(\mu) = -\gamma \sum_{i=1}^{N_k} (y_{i,k} - \mu_k(a_k))^2 \tag{4}$$

$$\Psi = \sum_{k=1}^{K} \mathcal{L}_k(\mu_{k-1}) - \mathcal{L}_k(\mu_k) \tag{5}$$

where μ_0 is an arbitrary prediction, ideally corresponding to a plausible ex ante prior belief, and γ is an arbitrary scaling parameter used to normalize units.

Note that because our approach elicits only posterior means, it is agnostic about the tools used to generate them. In particular, this process is independent of whether the experimenter generates their posterior using reduced-form or structural methods (Acemoglu, 2010; Garcia and Wantchekon, 2010; Ludwig et al., 2011; Duflo et al., 2012).

Proposition 7. *For all stages k, truthful reporting of μ_k maximizes subjective expected score $\mathbb{E}_{h_0|k}[\Psi]$. For all k, in equilibrium, an additional experiment increases the experimenter's expected score.*

Thus, to the extent that experimenters care about their reliability score, they should encourage replication. From their perspective, replication and out-of-sample tests increase their score in expectation. This may help alleviate the concerns that sometimes arise in the process of replication (Dercon et al., 2015; Bowser, 2015).

Using posteriors. The fact that posteriors are subjective means they should be used with some caution. When the number of implementations of a particular policy is relatively high, we are able to leverage the classic work of **??** and present an algorithmic decision rule that exploits the experimenter's posteriors over ranges of implementations in which she tends to be correct, and ignores her over ranges of implementations in which she tends to be wrong. Intuitively, this is achieved by starting out with a moderately high weight placed on the experimenter being correct, and increasing this weight when outcomes that agree with her predictions are reported, and lowering them when outcomes that disagree with her predictions are reported.

Suppose we have access to predictions $(\mu^k)_{k\in\{1,\cdots,K\}}$ corresponding to implementation opportunities in environments $(z_k)_{k\in\{1,\cdots,K\}}$ with measured treatment effects $\overline{p}^a_{z_k}$. Let $\eta_k \in \Delta(\{0,1,E\})$ denote the extent to which treatment 0, 1 or the treatment a^E_k recommended by the expert (that is, a such that $\mu^a_k - \mu^{1-a}_k > 0$) is implemented in the population. Let $\langle p_z, \eta\rangle = p^0_z\eta_0 + p^1_z\eta_1 + p^{a^E_k}_z\eta_E$ denote the average treatment effect given scaling rule η. We describe a rule of thumb for implementation η that exploit the expert's information when she is useful and ignore her otherwise. For $\theta \in \{0,1,E\}$, form regret measures

$$\mathcal{R}^\theta_k = w_\theta + \max_{k'\leq k} \sum_{t=k'}^{k-1} p^0_{z_t} - \langle p_{z_t}, \eta_t\rangle .$$

where w_θ is an arbitrary positive number (say 1), and $\langle p, \eta\rangle$ is the usual dot product. These regret measures capture the performance loss of allocation rule η against treatments 0, 1, and the recommendations of the expert, over the worst possible time range $[k', k]$. Define

implementation rule

$$\eta_\theta^k = \frac{\mathcal{R}_k^\theta}{\sum_{\theta' \in \{0,1,E\}} \mathcal{R}_k^{\theta'}}.$$

Proposition 8. *For all $\theta \in \{0, 1, E\}$, $k \in \{1, \cdots, N\}$, using η_θ^k above, $\mathcal{R}_k^\theta \leq \mathcal{O}\sqrt{N}$.*

In other words, within a loss of order $\sqrt{N}$, implementation rule η manages to follow the recommendations of the expert in implementations where she is correct, and uses the best possible fixed alternative when the expert is systematically mistaken. This implies that even though the recommendations of the expert are subjective, they may be exploited in robust ways.

Alternative reports. While we believe that expectations over mean outcomes in different external environments is a natural place to start, it is theoretically possible to extract other pieces of information. However, to actually implement the required elicitation mechanism may be impractical. Here we sketch briefly how incentives could be provided to truthfully report other simple moments of the experimenter's beliefs.

One may elicit both the experimenter's subjective mean μ_k and subjective standard deviation σ_k of $\overline{p}_z$ by using a modified score Ψ defined by

$$\forall k \in \{1, \cdots, K\}, \quad \mathcal{L}_k(\mu, \sigma) = -\gamma N_k \left[\frac{\left(\sum_{i=1}^{N_k} y_{i,k} - \mu_{z_k}^a \right)^2}{N_k \sigma_{z_k}^2} - 2 \log(\sigma_{z_k}) \right]$$

$$\Psi = \sum_{k=1}^{K} \mathcal{L}_k(\mu^{k-1}, \sigma^{k-1}) - \mathcal{L}_k(\mu^k, \sigma^k).$$

While asking experimenters to speculate about second moments is perhaps tricky, it seems useful to let experimenters express their confidence over the predictions they make in different environments.

One possible variant is to elicit whether the experimenter places subjective belief greater than ξ on the average treatment effect in environment z, $\overline{p}_z^1 - \overline{p}_z^0$, being greater than a

threshold ζ. This can be approximately achieved as follows. Fix some minimal sample size threshold $\underline{N}$, and consider reports $(m_k)_{k\in\{1,\cdots,K\}} \in \{0,1\}^K$ by the experimenter. Use the score Ψ defined by

$$\mathcal{L}_k(m) = -\mathbf{1}_{N_k \geq \underline{N}} \times (2m_{z_k} - 1) \left(\mathbf{1}_{\underline{y}_k^1 - \underline{y}_k^0 > \zeta} - \xi \right) N_k$$
$$\Psi = \sum_{k=1}^{K} \mathcal{L}_k(m_{k-1}) - \mathcal{L}_k(m_k).$$

6 Discussion

This chapter started with a discussion of statistical frameworks for investigating experimental design. We argue that using a maximin frameowrk captures best what experimenters are trying to do and provides useful insghts into experimental design. Finally, we propose a system to encourage experimenters to speculate about treatment effects in environments other than the one studied in a particular implementation.

Here, we attempt to discuss some qualitative aspects of a system for encouraging informed speculation about external validity. We emphasize that our discussion here is extremely preliminary, but we hope it can be part of a useful dialogue.

The technical details of our proposal to encourage informed speculation about external validity do not depend on the environments for which experimenters might choose to make predictions. Yet, in order for these predictions to be of some use, the set of environments would need to be constrained since the list of possible directions for extension is infinite. It makes sense therefore to try to build some consensus on what the most important directions for extrapolation should be.

Our discussion of these environments focuses on those likely to be of interest to a broad range of development economists. Other fields of economics, let alone medical applications, are likely to have somewhat different concerns, and would want to encourage speculation in other, particular, environments.

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.

Spill-over effects, which include what are somewhat misleadingly called general equilibrium effects, may be either positive or negative. To make this concrete, consider an intervention that gives scholarships for top performing students in local schools to attend provincial schools. In an experimental intervention, this 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, fewer students from each locality would qualify, possibly diminishing the effect on the original locality. On the other hand, higher caliber students would, on average, qualify for the scholarships, increasing the overall human capital of the province and country. To the extent that there are positive externalities to raises in human capital (as have been observed on the development path of many countries), this may increase the overall effect of the program when implemented at a national level.

Methods for designing experiments that at least partially capture these effects are now quite standard (Miguel and Kremer, 2004). However these methods rely on the assumption that spill-overs are mainly local, which may be the right assumption in some contexts but not others. It would be useful to get researchers to (a) try to incorporate such design elements wherever possible (b) explain why they did not adopt a method that allows them to estimate spillovers (for example, is it unlikely that there are in any) and (c) speculate, based on their understanding of the program and the relevant geographies, about what the range of effects at scale may be.

In contrast, informed speculation about implementing agencies is likely to be dicier. At the simplest level, three environments seem relevant: implementations by other researchers,

implementations by NGOs or international agencies, and implementations by provincial or country governments. However, in order to make these estimates meaningful, the experimenter should 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. it should be possible for the experimenter to highlight the specific aspects of the intervention that may or may not make it transferable to a different setting.

Finally, to the extent that there are positive or negative interaction effects between the scope of implementation and the implementing agency, an experimenter may wish to specify this.

Fixed Aspects of Individuals. Development economists have been particularly concerned with some specific attributes of the individuals in their study. Those that have garnered the most attention are gender, and religious and ethnic minorities. While these attributes would naturally be emphasized in any speculation about future results, differences in treatment effects across these groups are also likely to be directly measured in an experiment. Moreover, as many interventions are meant to explicitly address iniquities suffered by women and minorities, it may not make sense to speculate about the effects of such an intervention on men or an ethnic majority.

Thus, while it is easy to identify fixed attributes, coming up with a common set that should be part of any speculation about external environments may be more difficult, as these are likely to vary across experimental treatments. For example, an experimental examination of fertilizer might naturally include speculation about how that intervention might affect farmers of different crops. However, such speculation over these different types of farmers would be particularly uninformative if the experiment in question measured the impact of commitment-savings products on slum-dwellers.

Mutable Factors. Some of the issues with eliciting external predictions over fixed characteristics is that they are often a proxy for theoretically-relevant differences in mutable characteristics, such as beliefs, values, or credit constraints. As the goal of many policy interventions is to affect these underlying, mutable, characteristics, it may be easier both to speculate about the consequences of such changes. Of the many mutable characteristics of individuals, beliefs and credit constraints seem to be both focal to the take up of an intervention, and addressable through standard policy tools.

Additional Considerations

There are, no doubt, a number of practical details that would need to be sorted out, and challenges overcome, to make a system even remotely similar to the one we propose a reality. Even defining a standard set of groupings within each of the environments we suggest above would be challenging. At this point, we are not in a position to make a useful assessment of these challenges and practical details. Instead, we conclude with two points about the additional benefits that might accrue due to such a system.

While in most cases an experimenter's posterior will be close to the coefficients reported in her paper, an informed speculation section can allow the experimenter to present evidence or just her best guess based on casual observation, field experience, or unreported specifications that are outside the scope of the particular research question they are addressing in a rigorous way. In particular, this could reduce the incentives to present inductive (or data-mined) hypotheses as part of the main research, as 1) such hypotheses will now have a designated place in reporting research, and 2) there will be incentives to appropriately hedge such hypotheses.

As an example of the above, Deaton (2010, pp. 441–442) issues the following critique:

> [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, i.e., one that is unrelated to some underlying or guiding mechanism, is both unlikely and unlikely [sic.] to be persuasive.

Our proposal could help with such criticisms in three ways. First, it would establish a place within research where such speculation is both expected and encouraged. Second, by attaching 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 increases, at least slightly, the credibility of the experimenter's preferred underlying mechanism.

Finally, perhaps the most important side effect of asking experimenters to speculate about external validity is that it will create incentives to design experiments in a way that best allows them to address external questions. To address scalability, experimenters may structure more local pilot studies to allow easy comparison with their broader main experiment. To address immutable characteristics of individuals, experimenters may be more likely to include, and stratify sampling on, various subgroups, particularly ethnic minorities and women. To address mutable characteristics, such as beliefs or credit constraints, experimenters may elicit the former using the selective trial techniques discussed in Section 4.3. While these benefits are speculative, it is our belief that creating a structured framework to speculate about external validity is an important in promoting a successful ecosystem for social science field experiments, and a complement to many other aspects of experimentation.

References

Acemoglu, Daron, "Theory, General Equilibrium and Political Economy in Development Economics," 2010. NBER Working Paper Series # 15944. 5.2

Aghion, Philippe, Patrick Bolton, Christopher Harris, and Bruno Jullien, "Optimal learning by experimentation," *The review of economic studies*, 1991, *58* (4), 621–654. 1

Ashraf, Nava, James Berry, and Jesse M Shapiro, "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia," *American Economic Review*, 2010, *100* (5), 2383–2413. 4.3

Banerjee, Abhijit, "A Simple Model of Herd Behavior," *The Quarterly Journal of Economics*, August 1992, *107* (3), 797–817. 1

_ , Sylvain Chassang, Sergio Montero, and Erik Snowberg, "A Theory of Experimenters," 2015. Princeton University, *mimeo.* 1, 2, 3.1, 3.2, 3.2, 4.1, 4.1, 4.2, 3, 4, 4.2

Bellman, Richard, "A Problem in the Sequential Design of Experiments," *Sankhyā: The Indian Journal of Statistics*, April 1956, *16* (3/4), 221–229. 1

Bergemann, Dirk and Juuso Välimäki, "Learning and Strategic Pricing," *Econometrica: Journal of the Econometric Society*, September 1996, *64* (5), 1125–1149. 1

_ and _ , "Information Acquisition and Efficient Mechanism Design," *Econometrica*, 2002, *70* (3), 1007–1033. 1

_ and _ , "Bandit Problems," 2006. Cowles Foundation discussion paper. 1

Berry, James, Greg Fischer, and Raymond Guiteras, "Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana," 2012. Cornell University, *mimeo.* 4.3

Bewley, Truman F, "Knightian uncertainty," *ECONOMETRIC SOCIETY MONOGRAPHS*, 1998, *29*, 71–81. 3.2

Blackwell, David, "Comparison of Experiments," in "Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability," Vol. 1 1951, pp. 93–102. 3.2, 1

Bowser, William H., "The Long and Short of Returns to Public Investments in Fifteen Ethiopian Villages," 2015. International Initiative for Impact Evaluation, *mimeo.* 5.2

Chaloner, Kathryn and Isabella Verdinelli, "Bayesian Experimental design: A Review," *Statistical Science*, August 1995, *10* (3), 273–304. 1

Chassang, Sylvain, "Calibrated Incentive Contracts," *Econometrica*, 2013, *81* (5), 1935–1971. 5.2

_ **, 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. 1, 4.3, 4.3, 5

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. 4.3, 4.3

De Finetti, Bruno, "La Prévision: Ses lois Logiques, ses Sources Subjectives," *Annales de l'institut Henri Poincaré*, 1937, *7* (1), 1–68. 2

Deaton, Angus, "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, June 2010, *48* (2), 424–455. 6

Dercon, Stefan, Daniel O. Gilligan, John Hoddinott, and Tassew Woldehanna, "The Impact of Agricultural Extension and Roads on Poverty and Consumption Growth in Fifteen Ethiopian Villages: Response to William Bowser," 2015. International Initiative for Impact Evaluation, *mimeo.* 5.2

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. 6

_ **, Rema Hanna, and Stephen P. Ryan**, "Incentives Work: Getting Teachers to Come to School," *The American Economic Review*, 2012, *102* (4), 1241–1278. 5.2

Dupas, Pascaline, "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence From a Field Experiment," *Econometrica*, 2014, *82* (1), 197–228. 4.3

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

Ellsberg, Daniel, "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics*, 1961, *75* (4), 643–669. 3.2

Fisher, Ronald Aylmer, "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 1926, *33*, 503–513. 1

_ , *The Design of Experiments.*, Edinburgh and London: Oliver & Boyd, 1935. 1

Garcia, Fernando Martel and Leonard Wantchekon, "Theory, External Validity, and Experimental Inference: Some Conjectures," *The Annals of the American Academy of Political and Social Science*, 2010, *628* (1), 132–147. 5.2

Gilboa, Itzhak and David Schmeidler, "Maxmin Expected Utility with a Non-Unique Prior," *Journal of Mathematical Economics*, 1989, *18* (2), 141–153. 1, 3.2, 3.2

_ , Andrew Postlewaite, and David Schmeidler, "Is it always rational to satisfy Savage's axioms?," *Economics and Philosophy*, 2009, *25* (03), 285–296. 3.2

Gittins, John C., "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1979, *41* (2), 148–177. 1

Grossman, Sanford J and Joseph E Stiglitz, "On the Impossibility of Informationally Efficient Markets," *The American Economic Review*, June 1980, *70* (3), 393–408. 1

Heckman, James J. and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, May 2005, *73* (3), 669–738. 4.3

Karlan, Dean S. and Jonathan Zinman, "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," *Econometrica*, 2009, *77* (6), 1993–2008. 4.3

Kasy, Maximilian, "Why Experimenters Should not Randomize, and What they Should do Instead," 2013. Harvard University, *mimeo.* 3.1, 4.1, 7

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji, "A Smooth Model of Decision Making Under Ambiguity," *Econometrica*, 2005, *73* (6), 1849–1892. 1, 3.2

Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan, "Mechanism Experiments and Policy Evaluations," *Journal of Economic Perspectives*, 2011, *25* (3), 17–38. 5.2

Maynard, G.D., "Statistical Study of Anti-Typhoid Inoculation," *Biometrika*, March 1909, *6* (4), 366–375. 1

Miguel, Edward and Michael Kremer, "Worms: identifying impacts on education and health in the presence of treatment externalities," *Econometrica*, 2004, *72* (1), 159–217. 6

Milgrom, Paul R., "Rational Expectations, Information Acquisition, and Competitive Bidding," *Econometrica*, July 1981, *89* (4), 921–943. 1

Olszewski, Wojciech and Marcin Peski, "The principal-agent approach to testing experts," *American economic Journal: microeconomics*, 2011, *3* (2), 89–113. 5.2

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

Robbins, Herbert, "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, September 1952, *58* (5), 527–535. 1

Rothschild, Michael, "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 1974, *9* (2), 185–202. 1

Savage, Leonard J, *The Foundations of Statistics*, Courier Corporation, 1954. 3.2

Schmeidler, David, "Subjective Probability and Expected Utility without Additivity," *Econometrica*, July 1989, *57* (3), 571–587. 1, 3.2

Simpson, R.J.S. and Karl Pearson, "Report on Certain Enteric Fever Inoculation Statistics," *The British Medical Journal*, 1904, *2* (2288), 1243–1246. 1

Sims, Christopher A., "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690. 4.2

Vapnik, Vladimir, *The nature of statistical learning theory*, Springer Science & Business Media, 2000. 6