The Dozen Things Experimental Economists Should Do (More of)

Eszter Czibor*, David Jimenez-Gomez†, and John A. List‡

February 12, 2019 §

Abstract

What was once broadly viewed as an impossibility – learning from experimental data in economics – has now become commonplace. Governmental bodies, think tanks, and corporations around the world employ teams of experimental researchers to answer their most pressing questions. For their part, in the past two decades academics have begun to more actively partner with organizations to generate data via field experimentation. While this revolution in evidence-based approaches has served to deepen the economic science, recently a credibility crisis has caused even the most

*Department of Economics, University of Chicago, 1126 E. 59th, Chicago, IL 60637, USA; E-mail: eczibor@uchicago.edu

†Department of Economics, University of Alicante, Spain; Email: davidjimenezgomez@ua.es

‡Department of Economics, University of Chicago, 1126 E. 59th, Chicago, IL 60637, USA; E-mail: jlist@uchicago.edu

§This paper is based on a plenary talk given by John List at the Economics Science Association in Tucson, Arizona, 2016, titled "The Dozen Things I Wish Experimental Economists Did (More Of)". We are grateful for the insightful suggestions provided by Jonathan Davis, Claire Mackevicius, Alicia Marguerie, David Novgorodsky, Julie Pernaudet and Daniel Tannenbaum, and for great comments by participants at the University of Alicante research seminar, at the Barcelona GSE Summer Forum's Workshop on External Validity, and at the 4th Workshop on Experimental Economics for the Environment at WWU Münster. We thank Eric Karsten, Ariel Listo and Haruka Uchida for excellent research assistance.

ardent experimental proponents to pause. This study takes a step back from the burgeoning experimental literature and introduces 12 actions that might help to alleviate this credibility crisis and raise experimental economics to an even higher level. In this way, we view our "12 action wish list" as discussion points to enrich the field.

JEL Classification: A11, C91, C93

"There is a property common to almost all the moral sciences, and by which they are distinguished from many of the physical... that it is seldom in our power to make experiments in them", Mill (1836, p.124).

"Unfortunately, we can seldom test particular predictions in the social sciences by experiments explicitly designed to eliminate what are judged to be the most important disturbing influences. Generally, we must rely on evidence cast up by the 'experiments' that happen to occur", Friedman (1953, p.10).

"Economists cannot make use of controlled experiments to settle their differences: they have to appeal to historical evidence", Robinson (1977, p.1319).

"The economic world is extremely complicated. There are millions of people and firms, thousands of prices and industries. One possible way of figuring out economic laws in such a setting is by controlled experiments... like those done by chemists, physicists, and biologists... Economists have no such luxury when testing economic laws. They cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe", Samuelson and Nordhaus (1985, p.8)

Introduction

The give and take between theory and data in the natural sciences is so ingrained in modern thought that an integral part of the scientific method – that theories must be tested

against experimental evidence – is now second nature. This fact, of course, was not lost on the icons of economics, many of whom felt compelled to express their anguish by comparing empirical approaches across the social and natural sciences. The common thread in the epigraph musings is that if economists desire to do experimentation they should choose another practice, and if they want to engage in empirical economics, they should start looking for available naturally occurring data. This is presumably because the writers believed that it was impossible to collect/learn from experimental data in economics. These general feelings were shared ubiquitously throughout the 19$^{\text{th}}$ and 20$^{\text{th}}$ centuries, as extracting knowledge from historical data and personal introspection represented the primary source, and indeed in most cases the sole source, of new empirical knowledge in economics.

The economic landscape is changing. In the past several decades constructing new approaches to generate data have opened up several avenues for a fresh approach to understanding the economic relationship of theory and data. Whether by lab or by field, the popularity of experiments in economics has steadily increased, in large part due to the advantages they offer in identification, control, statistical inference, and interpretability. Properly constructed experiments take the analyst beyond measurement, into the "whys" of the causal relationship. It is often within these "whys" where the deep theoretical understandings or the key policy takeaways reside (see, e.g., List (2004b) on using field experiments to understand the nature of discrimination observed in markets).

While many would consider using randomization in the lab and the field as an unequivocal success in moving economics to a new scientific level, recently critics in the broader social sciences have called for the movement to proceed more cautiously. As Maniadis et al. (2017) point out, an active debate has surfaced that claims there is a "credibility crisis" in several scientific disciplines, including psychology (Nosek et al. 2012), management (Bettis 2012), and several branches of the biological and human sciences (e.g., Jennions and Møller (2003); Ioannidis (2005)). While the crises take many forms, one com-

mon widespread concern revolves around reproducibility, with the rate of 'false positives' representing a particular concern.

This literature motivated us to step back from the evidence-based movement and ask a simple question: if we could gather social scientists in a room and had the goal of enhancing knowledge discovery, what advice would we give to experimental researchers? This thought experiment yields a wish list of 12 things that we hope experimental economists will do more of in the future. We group our list of 12 recommendations into three bins. We begin with the decision concerning what data to acquire to ensure the generalizability of our results. We proceed to discuss best practices to generate informative and credible evidence via experimentation. We conclude with ways to interpret, build on, and scale the initial experimental evidence to make it useful and relevant for practitioners. We represent these bins by three broad questions below.

(1) What data should we acquire? We begin by calling researchers to carefully consider the generalizability of their findings not only after the analysis stage, but already when making the data acquisition choice. To facilitate this approach, we highlight systematic threats to generalizability in experiments. These considerations prompt us to advocate for running more field experiments, especially natural field experiments. We do so not only because natural field experiments are relatively new compared to many other empirical approaches, and therefore much ground is untrodden, but also because they provide a unique mix of desirable features – randomization and realism – that other approaches have difficulty combining by their very nature. This bin concludes with a call for using lab and field experiments (as well as naturally-occurring data) as *complements* in the evidence generation process. This is important because they each provide different parameters of interest (see Al-Ubaydli and List (2015)) and address the aforementioned threats differently.

(2) How should we generate data and interpret information from experiments? This second bin collects considerations for the design and analysis of experiments to make our

results more informative and credible. We first highlight the dangers of over-reliance on p-values for inference, and discuss alternatives. Complementing that interpretational question is a discussion of proper replication. Replication was a leg of the Fisher (1935) experimental tripod and represents a signature issue within the credibility revolution that we simply echo here, alongside recommendations for incentivizing replications. Our second bin also includes four design elements that critically determine what, and how much, we can learn from the data generating process: adequately considering statistical power in the design phase; adjusting for multiple hypothesis testing (a common reason for false positives) not only in our data analysis but also in our designs; using blocked randomization to increase power and ensure balance; and understanding heterogeneity through within-subject variation when necessary and appropriate.

(3) How can we produce evidence that is relevant for policy making? Our last bin revolves around how experimentalists can most usefully assist policy makers. Perhaps surprisingly, this discussion begins with advocating for a deeper use of theory to motivate designs by going beyond typical A/B tests. This approach importantly allows the analyst to determine the "whys" behind a result, leading to more effective policies. In addition, it helps to maintain fidelity of the program when policymakers scale the intervention. Scaling needs science in and of itself, and that science should be considered early in the experimental design phase. Complementing this discussion is a plea for experimentalists to go beyond measurement of short-run substitution effects to focus also on long run effects– these are behavioral effects that policymakers find of great import but are often not provided in the literature.

In order to help the reader to better navigate the paper, we include here a summarized list of our twelve suggestions.

1. **Appropriately consider generalizability, across the lab and the field.** We provide a framework for assessing the generalizability of experimental results, i.e

whether the result continues to hold when some variable in the experiment is changed. We identify threats to generalizability in the following areas: the characteristics of the experiment itself, subjects' participation and compliance decisions, and the representativeness of the sample.

2. **Do more field experiments, especially natural field experiments**, because they are uniquely suited to deal with many of the threats to generalizability outlined in Section 1, and because they provide unique control over the participation decision.

3. **Use lab and field experiments as complementary approaches** in the production of scientific knowledge. We recommend that researchers choose the type of experiment that is the best fit for the scientific question at hand, taking into account the level of control on the task, the generalizability, and the costs, and also combine different types of experiments for the best results.

4. **For proper inference, go beyond p-values.** This includes recognizing the importance of statistical power to avoid false negatives and effect inflation. We suggest adopting a Bayesian framework of inference that explicitly considers the priors (of researchers and/or the research community) about the studied phenomena.

5. **Replicate early and often**. Replication should be an integral part of the experimental process, but in reality it remains rare. We show the value of replication in a Bayesian framework, and discuss possible ways to incentivize researchers to conduct replication studies.

6. **Consider statistical power in the design phase.** Rather than ex post power calculations, we advocate for taking power seriously ex ante. To assist this process, we provide an overview of practical issues related to sample size calculations.

7. **Adjust for multiple hypothesis testing, in power tests and in data analysis.** The practice of simultaneously conducting multiple comparisons is widespread in the experimental literature, and can lead to high false positive rates. We discuss different methods to deal with this challenge, and focus in detail on controlling the family-wise error rate.

8. **Use blocked randomization to increase power and credibility.** When baseline characteristics of the participants are observable, researchers should utilize this information when they assign subjects to treatment through blocked randomization. This practice can increase the power of the study – and allow the researchers to signal which dimensions of heterogeneity they find ex ante important. We also discuss methods other than randomization for treatment assignment.

9. **Use within-subject designs when appropriate.** In cases when potential biases from learning or sensitization do not pose a serious threat, researchers should consider using a within-subject design (in which the same subject is exposed to multiple treatments sequentially), as it often yields greater statistical power than between-subject designs, and may help reveal heterogeneous treatment effects.

10. **Go beyond A/B testing by using theoretically-guided designs.** Incorporating economic theory into the design of experiments allows researchers to explore the underlying mechanisms that cause an effect, to estimate structural parameters, to conduct welfare analysis, and to better capture general equilibrium and spillover effects.

11. **Focus on the long run, not just on the short run.** Measuring the long-run effect of treatments is crucial for ROI calculations, and for adequately estimating welfare and general equilibrium effects.

12. **Understand the science of scaling ex ante and ex post.** We argue that scaling should be treated as a scientific problem in its own right. We provide a framework that incorporates the ideas expressed throughout this paper, allowing researchers to "backward induct" and address potential threats to scalability already in the design of experiments.

We contend that many of the questions discussed in this paper are not restricted to experiments; the issues of generalizability, causal inference, replication, power, correcting for multiple hypothesis testing, the use of theory, measuring long-term effects, etc. are relevant for applied economics research more generally. Moreover, despite our usage of the term "experimentalists", we do not view the experimental method as confined to a subset of the profession. Rather, we believe experiments may serve as a helpful tool for economists who are active in *any* field, given the right circumstances. We therefore hope that our paper can be informative not only for those scholars who are already engaged in conducting experiments, but for any economist who has ever considered running one.

The remainder of our study proceeds as follows. The next section presents preliminaries and sets the stage for the development of our three bins. We then describe our views on what data to acquire, how to generate data and create useful information via experimentation, and how to interpret, build on, and scale the initial experimental evidence. This discussion yields our dozen ideas that we hope experimental economists will do more of in the future. Throughout the paper, we point our readers to inspiring examples of experiments that engage in the practices we advocate for. We conclude with summary thoughts.

Preliminaries: experiments, estimation and randomization

This section offers an overview of the most relevant concepts to be applied throughout the paper (readers familiar with the inferential basics of economic experiments may wish to skip this overview, and start directly at Section 1). We begin by defining our subject of interest. In principle, we consider an **experiment** broadly: a study that generates primary, or original, data in a controlled environment. This inclusive definition permits the original studies of early experimentalists to be classified as experiments as well as important studies that exploit controlled environments but use non-experimental variation as the primary means of identification.[1] While we view experiments broadly, for the purposes of this study, we take a narrower definition of experiment by considering only those studies where researchers generate primary data by using randomization to identify a causal relationship.[2] Accordingly, there is a clear difference between these type of data and data obtained from so-called **natural experiments**, where subjects are randomly allocated to different treatment groups by a process outside of the researcher's control (such as the draft lottery in Angrist 1990); or from **quasi-experiments**, in which subjects are not randomly assigned to treatments (Greenstone and Gayer 2009). The interested reader can find more on the different types of experiments in Shadish et al. (2002).

In the following, we present a framework intended to guide our discussion of experiments, formalizing the most important concepts used in the paper. Individual i has co-

[1] Examples include studies that, instead of comparing outcomes between a treated and a control group, make comparisons along pre-existing traits of their subjects, such as their gender, age, religion, occupation, etc. Consider, for instance, Koudstaal et al. (2015) who study differences in risk attitudes among entrepreneurs, managers and employees.

[2] Our definition is in the spirit of Shadish et al. (2002), who define an experiment as "a study in which an intervention is deliberately introduced to observe its effects", and a randomized experiment, in addition, must be such that "units are assigned to receive the treatment or an alternative condition by a random process".

variates x_i. The experiment has characteristics ω, where ω includes the subject population (university students, CEOs, etc.), context (artificial vs. natural), time horizon (long vs. short), and other characteristics. The experiment consists of the following stages:

- Let p_i be an indicator variable for the **participation decision** such that $p_i = 1$ if subject i chooses to participate in the experiment, and $p_i = 0$ otherwise.
- Let z_i denote **assignment to treatment**. For example, $z_i = 1$ if student i is assigned to a small class size (z_i will be random in the experiments we discuss).
- Let d_i be the **treatment status**, which is the treatment individual i actually receives (e.g. $d_i = 1$ if student i actually attends a small class). Note that it is possible that z_i and d_i are different.[3]
- Let y_{i1} be the **outcome of interest** (e.g. the child's test scores) when treatment status is $d_i = 1$, and y_{i0} when treatment status is $d_i = 0$.[4]

We follow the **potential outcomes framework**, in which an individual i has outcome y_{i1} in the treated group and y_{i0} in the control group. Ideally, when conducting an experiment, researchers would like to measure individual treatment effects for each individual i, $y_{i1} - y_{i0}$, which is the difference in outcomes for individual i being in the treated versus the control group. In practice, of course, they cannot observe both of these outcomes; instead they can only observe individual outcomes in one of the treated states, and the *counterfactual* outcome in the other state remains unobserved. Instead of individual treatment effects, researchers therefore usually consider the **Average Treatment Effect (ATE)**,

[3]There are two possible cases: (1) The subject is assigned to a treatment z, such as a voucher to enroll in training, that is of a different nature than the treatment status, which is whether the subject actually enrolled in treatment. In that case, $Z \neq D$. (2) Alternatively, the subject is assigned to a treatment z, which is already one of the potential treatment statuses. For example, the subject is assigned to a training course $z = 1$ or not $z = 0$. Subjects can still opt in ($d = 1$) or out ($d = 0$) of the training course, and in this case $Z = D$.

[4]Our framework follows the tradition of Rubin (1974), which can traced back to the work of Jerzy Neyman; see also Freedman (2006).

given by $\tau^* = \mathbb{E}[y_{i1} - y_{i0}]$. The ATE measures the average difference in the outcomes for the population. The ATE τ^* is not directly observable either; instead researchers estimate τ, defined as:

$$\tau = \mathbb{E}[y_{i1}|d_i = 1] - \mathbb{E}[y_{i0}|d_i = 0].$$

Estimate τ measures the difference between the average effect of the treatment on those who were treated and the baseline average outcome of those who were not treated. As it will become clear below, when: 1) d is randomly assigned, 2) subjects do not opt in or out of their assigned treatments, and 3) potential outcomes of an individual are unrelated to the treatment status of any other individual, then $\tau = \tau^*$.[5] Note that the ATE does not allow researchers to estimate the percentiles of the distribution of treatment effects, or other moments such as the variance (we discuss estimating heterogeneous treatment effects in Section 9) and, unlike measures based on percentiles such as the median, the ATE is sensitive to outliers, observations whose value greatly differs from the rest (Deaton and Cartwright 2018). Note also that the "experiment population" is not necessarily a random sample of the entire population and may be selected according to observables; in that case, we only learn the effect of the treatment on the particular sub-population from which the sample is drawn (Duflo et al. 2007), an issue we discuss in detail in Section 1.

In the absence of randomization, researchers estimate

$$\tau = \mathbb{E}[y_{i1}|d_i = 1] - \mathbb{E}[y_{i0}|d_i = 0] = \underbrace{\mathbb{E}[y_{i1} - y_{i0}|d_i = 1]}_{\text{ATE on the treated}} + \underbrace{\mathbb{E}[y_{i0}|d_i = 1] - \mathbb{E}[y_{i0}|d_i = 0]}_{\text{selection bias}}$$

A non-zero **selection bias** term in the previous equation indicates that those who select into treatment are different in the untreated state from those who do not sort into

[5] The third assumption is the "Stable Unit Treatment Value Assumption" (Angrist et al. 1996; Duflo et al. 2007), which assumes away any spillover effects.

treatment. This happens, for example, if smokers who are more motivated to quit are more likely to enroll in a smoking cessation treatment than those who are unmotivated: in such a case, we end up with program participants who are inherently different (more motivated) than those who did not take up the program, leading to a biased (in our case, overoptimistic) estimate of the program's effect on quitting. In order to rule out selection bias, it is necessary to make certain assumptions, such as the Conditional Independence Assumption (Rosenbaum and Rubin 1983):

$$\{y_{i0}, y_{i1}\} \perp\!\!\!\perp d_i | x_i,$$

which claims that the outcome in each state and the assignment to treatment for a given individual are independent conditional on the observable covariates. Intuitively, the Conditional Independence Assumption means that conditional on the observables x_i, the assignment to treatment is as good as random, and it implies that $\mathbb{E}[y_{i1}|x_i, d_i = 1] - \mathbb{E}[y_{i0}|x_i, d_i = 0] = \mathbb{E}[y_{i1} - y_{i0}|x_i]$, and therefore that $\tau = \tau^*$.

Crucially, random assignment to treatment automatically implies the Conditional Independence Assumption and hence solves the issue of selection bias (Duflo et al. 2007). As such, the most important reason why researchers (not just economists) use randomization is because it allows causal inference under potentially weaker assumptions than alternative methods.[6] Randomization serves as a novel instrumental variable (IV), balancing unobservables across control and treatment groups (Al-Ubaydli and List 2013).[7]

Studies based on random assignment also have the advantage of being easily replicable,

[6] A recent study suggests that in the presence of non-i.i.d. errors, IV estimates can have lower power than usually assumed, and a reassessment of published work suggests that statistically significant IV results depend heavily on a few observations, and provide little statistical evidence of a bias in ordinary least squared (OLS) estimates (Young 2017). These issues typically do not arise in a well-designed randomized experiment.

[7] For a discussion of some popular non-experimental methods, and their comparison to experiments, see Duflo et al. (2007). For a comprehensive discussion of the problems of randomization, see Deaton and Cartwright (2018).

in contrast to methods that rely on baseline covariates to assign treatments without randomization.[8] Random assignment can also deal with three challenges related to practical implementation and feasibility: first, it prevents the experimenter from allocating subjects to treatment and control in ways that would bias the results (for example, politicians assigning their constituents to a "schooling" treatment, or physicians assigning patients with higher perceived need to treatment). Second, it provides a credible way to measure treatment effects because it allows for a straightforward calculation of mean differences between the treatment groups where researchers have little leeway. Third, randomization is crucial in instances when fairness and transparency are a concern, because it insures against favor/discrimination towards particular groups.[9]

Though applied economists typically use conventional, sampling-based tests to analyze data from experiments, random assignment to treatment also allows for the construction of exact tests that do not rely on assumptions about the sample size or the error structure (Young 2019). In essence, randomization-based inference treats subjects' potential outcomes as fixed, and considers their assignment to treatment as random – an approach better fit for analyzing experimental data than sampling-based inference that assumes that treatment assignment is fixed, outcomes are random, and subjects are drawn from a much larger population (Athey and Imbens 2017a). A drawback of randomization-based inference is that it provides an exact test of a *sharp null*: one that specifies a precise treatment effect for each participant (Young 2019). Rather than testing whether the average treatment effect was zero, this approach only allows us to test the null hypothesis that the treatment had no effect on any participant at all – a null that Young (2019) considers stringent but not unreasonable.

[8] We return to the topics of replicability in Section 5 and optimization-based methods (e.g. Kasy 2016) in Section 8.

[9] Note that methods other than randomized experiments can achieve this goal too, see Deaton and Cartwright (2018); Kasy (2016); Banerjee et al. (2017b). We discuss scaling up further in Section 12.

To conclude, while randomization does not solve all (theoretical or practical) problems related to causal inference, when proposing alternatives to randomization in experiments, researchers should be very precise about the exact details of the alternative they propose, or else they run the risk of underestimating the value of experimentation (Senn 2013).

Throughout our paper, we follow the taxonomy for experiments developed by Harrison and List (2004), who identify four general categories, as summarized in Table 1.[10] **Laboratory experiments** study university students as subjects in an artificial environment i.e. the lab. For example, Goeree and Holt (2001) have university students play a series of games, showing how the predictive power of the Nash equilibrium is not robust to (supposedly innocuous) changes in payoffs in those games. **Artefactual Field Experiments (AFE)**, also known as *lab-in-the-field*, share most of the characteristics of lab experiments (such as having an artificial environment), but use the relevant population of interest as subjects. For example, Levitt et al. (2009) observed chess players at two international open tournaments to gather data on strategic behavior on some well-known games. **Framed Field Experiments (FFE)**, like AFE, use the relevant population as subjects, but take place in a natural environment, such as the market, school, hospital, etc. For example, Gosnell et al. (2017) incentivized airline captains to improve efficiency and save fuel (via performance information, personal targets, and prosocial incentives), and the pilots were aware that an experiment was taking place. Note that all three types of experiments described above are *overt*: subjects are aware of being part of an experiment.

In contrast, **Natural Field Experiments (NFE)** are *covert*: they study the relevant population in a natural setting and, crucially, subjects are not aware of being part of an experiment, setting NFE apart from the other types of experiments, as we discuss further below. For example, Hallsworth et al. (2015) randomized the letters sent to individuals who had debt obligations with the government in UK (the treatment group had an extra

[10] See also Karahanna et al. (2018) for a related discussion on the online variants of experiment types.

	Lab	**AFE**	**FFE**	**NFE**
Population we study	U	S	S	S
Environment	A	A	N	N
Type of awareness	O	O	O	C
Who do we observe?	$p_i = 1$	$p_i = 1$	$p_i = 1$	All

Table 1: Summary of the characteristics of each type of experiment. Population can be *U*niversity students or the *S*pecial population of interest. The environment can be *A*rtificial or *N*atural. The experiment can either be *O*vert or *C*overt.

sentence that informed them that refusal to pay would be considered as an active choice). In this case, subjects belonged to the relevant population and were in a natural context; moreover they were not aware of being part of an experiment.[11]

In sum, we can define the relevant estimates from lab, AFE, FFE and NFE as:

$$\tau^{lab} = \mathbb{E}[\tau | i \in U, e = A, t = O, p = 1],$$
$$\tau^{AFE} = \mathbb{E}[\tau | i \in S, e = A, t = O, p = 1],$$
$$\tau^{FFE} = \mathbb{E}[\tau | i \in S, e = N, t = O, p = 1],$$
$$\tau^{NFE} = \mathbb{E}[\tau | i \in S, e = N, t = C],$$

where U and S refer to students vs. a special population, the environment e can be artificial (A) or natural (N), the type t of experiment can be overt (O) or covert (C), and p indicates the presence or absence of an active decision to participate in the experiment.

With these preliminaries mind, we turn to the dozen things we hope experimentalists do more of. While there is no inherent ordering by importance of our 12 ideas, we attempted to group the topics loosely by what data to generate, how to efficiently generate and interpret the data, and how to give the most informative advice to evidence-based policymakers.

[11] Randomized Controlled Trials (RCT) would fall under either the FFE or the NFE classification, depending mainly on whether subjects are aware of being part of an experiment or not.

1 Appropriately consider generalizability, across the lab and the field

When designing an experiment, researchers need to balance two key aspects that determine the value of their contribution to science and policy. One aspect is **correct statistical inference**, including internal validity (the property of being able to identify the parameters of interest in a given design) and informativeness (how much a result can change the prior of the scientific community). The second is **generalizability** (also known as external validity): whether a causal relationship continues to hold when subjects, context, location, or treatment details are modified (Shadish et al. 2002). In what follows, we use the term generalizability instead of external validity, following Harrison and List (2004). This section outlines a framework for discussing threats to generalizability, building on the basic ingredients introduced in the Preliminaries. In the two sections that follow, we then use this framework to evaluate the different types of experiments (laboratory, artefactual field, framed field and natural field experiments) as defined in the Preliminaries.

The question of generalizability has long been studied in the social sciences, but has been often obfuscated, especially in non-experimental research, by the more pressing problem of internal validity (Al-Ubaydli and List 2013; Deaton and Cartwright 2018). While internal validity is necessary for generalizability, it is not sufficient (Duflo et al. 2007). In economics, the "Lucas critique" (Lucas 1976) famously tackled the issue of generalizability, by arguing against econometric policy evaluations that failed to recognize that agents' behavior varies systematically with changes in policy.[12] More recently, a new literature on "generalizability theory" has grown within psychology and economics (Briggs and Wil-

[12] In particular, the Lucas critique censured using estimates from past data to forecast the effects of a new policy, because the behavior of the agents will change in response to the implementation of the new policy, invalidating those estimates (Ljungqvist 2008). The interested reader will also find Goodhart's Law and Campbell's Law as two social science contemporaries.

son 2007; Higgins and Thompson 2002; Al-Ubaydli and List 2013).[13] However, the topic of generalizability still requires more attention in our field: in a survey of RCTs conducted in developing countries and published in leading economics journals, Peters et al. (2018) found that most of the papers did not discuss potential threats to generalizability, and argued for the peer review process to explicitly consider design features that would be relevant for generalization.

In order to improve generalizability of research findings, it is useful to classify the potential threats to generalizability according to their causes. We have identified four potential threats to generalizability: interaction between treatment and other characteristics of the experiment, selective noncompliance, non-random selection into the experiment, and differences in populations.

Threat I: Characteristics of the experiment. Characteristics inherent to the experiment can inadvertently affect outcomes and thus complicate the interpretation of results. In all experiments, y will be affected by the elements of ω, such as scrutiny, stakes, the time horizon of the intervention, and the environment (artificial or natural) (Deaton and Cartwright 2018). As such, it may not be possible to generalize our estimates to settings where those parameters are different.[14] **Overt** experiments, in which subjects are aware of being part of an experiment (such as lab experiments, artefactual field experiments (AFE), and framed field experiments (FFE)) are particularly prone to this threat to generalizability. The high level of scrutiny present in overt experiments may induce "experimenter demand effects", such that subjects attempt to behave in the way they

[13] See Briesch et al. (2014) for an introductory article to generalizability theory. Vivalt (2017) used techniques from generalizability theory to perform a meta-analysis of 20 types of intervention in economic development, collected from 635 papers, and found that results are more heterogeneous than in other fields such as medicine. Within generalizability theory, there is also an intriguing approach that attempts to generalize by establishing networks of causality (Bareinboim and Pearl 2013).

[14] Notice that the definition of the average treatment effect for the different types of experiments ($\tau^{lab}, \tau^{ATE}, \tau^{FFE}$ and τ^{NFE}, presented in the Preliminaries) all depend on ω, the characteristics of the experiment.

believe the experimenter wants them to (Quidt et al. 2018). Additional threats include Hawthorne and John Henry effects.[15] Several experiments studying e.g. pro-social behavior have demonstrated that scrutiny can indeed affect participant behavior (Bandiera et al. 2005; List 2006b; Benz and Meier 2008; Alpizar et al. 2008).[16]

Threat II: Selective noncompliance. We define as noncompliance instances when subjects end up, either by omission or by commission, *receiving* a different treatment than what they were initially *assigned* to. Noncompliance is especially problematic when subjects actively change their treatment, e.g. because they derive higher utility from a different treatment than the one they were assigned to, causing what is known as a **selection problem** (Heckman 2010, see also Footnote 3).[17] Let Z be the set of assignments to treatment in the experiment, and D the set of treatment statuses in the experiment, so that $z_i \in Z$ and $d_i \in D$. In the most general framework, subject i is assigned to treatment z_i, and there is a selection function that determines which treatment status d_i the subject ends up with. For example, subject i has $d_i = \arg\max_{\hat{d}_i \in D} u(x_i, \omega, \hat{d}_i) - C(x_i, z_i, \hat{d}_i)$, where $u(x_i, \omega, d_i)$ is the subject's utility of being in treatment status d_i, and $C(x_i, z_i, d_i)$ is her cost of choosing d_i conditional on being assigned to z_i. In these cases, the researcher assigns z_i, and then the subject chooses d_i to maximize her utility net of switching costs.[18]

[15] The Hawthorne effect is defined by the Oxford English Dictionary as "an improvement in the performance of workers resulting from a change in their working conditions, and caused either by their response to innovation or by the feeling that they are being accorded some attention"; for a review and a re-analysis of the data from the original Hawthorne experiment, see Levitt and List (2011). The John Henry effect refers to subjects exerting greater effort because they treat the experiment like a competitive contest (Horton et al. 2011).

[16] Camerer (2015) argues that scrutiny is not likely to affect subject's behaviors, based on the fact that subjects cannot usually guess the purpose of the study (Lambdin and Shaffer 2009), or that people are also usually observed when making real-life economic decisions (Falk and Heckman 2009). However, we believe that the scrutiny in overt experiments is of a much higher degree than what subjects normally experience, and that it likely affects behavior *directly*, even if subjects cannot correctly guess the purpose of the study.

[17] We highly recommend the recent paper by Kowalski (2018), who argues that rather than considering it a nuisance, researchers could treat this type of selection as a useful source of information that (combined with certain assumptions) can speak to the external validity of their experiment.

[18] Note that whether subjects solve this maximization problem ex-ante (so that they sort into treatment groups) or ex-post (they switch treatment groups) can have consequences for estimation (Heckman 2010).

As a result, we may observe $z_i \neq d_i$ for some individuals. In the case of imperfect compliance, we have that $\tau^* = \mathbb{E}[y_{i1} - y_{i0}] \neq \mathbb{E}[y_{i1}|d_i = 1] - \mathbb{E}[y_{i0}|d_i = 0] = \tau$.

When researchers cannot obtain the ATE due to noncompliance, they can instead estimate the "Policy Relevant Treatment Effect" (which, in the case when z_i is uncorrelated with y_i, coincides with the "Intention to Treat Effect" (ITT), Heckman 2010). The ITT might be the relevant estimate in some situations, because it provides researchers with a measure of how much the intervention "converts" into outcomes, as it considers the difference in outcomes between those who were *initially* assigned to treatment and the control group, irrespective of whether they complied with their treatment assignment. Researchers can also estimate the "Local Average Treatment Effect" (LATE), Angrist and Imbens (1994):

$$\text{LATE}_{p=1} = \mathbb{E}[y_{i1} - y_{i0}|\omega^{FFE}, d_i(z_i = 1) = 1, d_i(z_i = 0) = 0, p_i = 1], \tag{1}$$

where p_i refers to the decision of participating in the experiment (see Preliminaries). The LATE measures the average treatment effect for individuals induced into treatment $d_i = 1$ by a change in z_i (Heckman 2010). Note, however, that the average treatment effect measured by the LATE is only valid for that particular subpopulation (the compliers), and might differ from the ATE for the whole population, limiting its generalizability.[19]

An extreme case of non-compliance would be **attrition**, in which subjects leave the experiment (and their outcomes are therefore no longer observable to the experimenter). While random attrition only reduces power, attrition that is not random can bias the results (Duflo et al. 2007), for example when those individuals who are the most motivated leave the experiment if they are not assigned to a certain treatment. The best approach to solving inference problems related to attrition is to design the experiment in a way that

[19]For a more detailed account of LATE, and the conditions for its use, see Angrist and Imbens (1994) and Heckman (2010).

allows researchers to track subjects even if they leave the experiment (for more details, see Duflo et al. 2007), or to conduct a natural field experiment.

Threat III: Non-random selection into the experiment. As we have seen in the Preliminaries, treatment effect estimates from lab experiments, AFE and FFE are only valid for those individuals who select into the experiment (those with $p_i = 1$). The ability of these experiments to identify parameters of interest thus depends on assumptions about individuals' decision to select into the experiment. When participation in the experiment is not random but instead is the result of a cost/benefit analysis by the subjects, **participation bias** may arise (Al-Ubaydli and List 2013; Slonim et al. 2013).

Recall that the parameter of interest is the ATE for the whole population: $\tau^* = \mathbb{E}[y_{1i} - y_{0i}]$. Overt experiments, however, provide the following estimate: $\mathbb{E}[y_{1i} - y_{0i}|p_i = 1]$. The ATE τ^* is given by: $\tau^* = \mathbb{P}[p_i = 1] \cdot \mathbb{E}[y_{1i} - y_{0i}|p_i = 1] + \mathbb{P}[p_i = 0] \cdot \mathbb{E}[y_{1i} - y_{0i}|p_i = 0]$. Because $\mathbb{P}[p_i = 0] = 1 - \mathbb{P}[p_i = 1]$, we can compute the participation bias:

$$\underbrace{\mathbb{E}[y_{1i} - y_{0i}|p_i = 1] - \mathbb{E}[y_{1i} - y_{0i}]}_{\text{participation bias}} = \mathbb{P}[p_i = 0] \times (\underbrace{\mathbb{E}[y_{1i} - y_{0i}|p_i = 1] - \mathbb{E}[y_{1i} - y_{0i}|p_i = 0]}_{\text{treatment specific selection bias}}). \tag{2}$$

In other words, participation bias is the product of the probability of not being in the experiment $\mathbb{P}[p_i = 0]$ and the Treatment Specific Selection Bias (which is analogous to the classical selection bias, except that the selection is with respect to participation in the experiment, Al-Ubaydli and List 2013).[20]

Because in general $\mathbb{P}[p_i = 0]$ is very large (usually close to 1), the bias in the estimate will be determined mainly by the Treatment Specific Selection Bias, a fact anticipated by Slonim et al. (2013). Participation bias does not present a problem in overt experiments when $\mathbb{E}[y_{1i} - y_{0i}|p_i = 1] \approx \mathbb{E}[y_{1i} - y_{0i}|p_i = 0]$. This happens when p_i is independent of y_i,

[20] When researchers' goal is to obtain the Intent-to-Treat (ITT) or the Average Treatment Effect on the Treated (ATT) estimates, participation bias presents less of a problem, in the sense that researchers are interested in estimating effects for those who choose to participate in the experiment anyway.

either because selection does not depend on x_i, or because selection depends on some subset of x_i which is in turn independent of y_i. The following condition for overt experiments guarantees that the Treatment Specific Selection Bias will be zero:[21]

$$\{y_{i0}, y_{i1}\} \perp\!\!\!\perp p_i | x_i \qquad \text{(Generalizability Independence Condition (GIC))}.$$

Note that participation bias can arise even if one is conducting a standard lab experiment and the effect we are looking for can reliably be found in university students, in an artificial environment, with low stakes and with scrutiny (so the first threat to generalizability is not a concern), and even if $z_i = d_i$ for all individuals (so the second threat to generalizability is not a concern either). Slonim et al. (2013) found that, from a population of roughly 900 university students, those who selected into lab experiments had less income, more leisure time, more interest in economics and were more pro-social in the dimension of volunteering, all of which are consistent with participation being the result of a cost/benefit decision. Moreover, risk averse individuals might be less likely to enroll in an experiment (Al-Ubaydli and List 2013; Heckman 2010). Participation bias may also arise in the field, because organizations who agree to collaborate with researchers in an experiment are usually exceptional (Banerjee et al. 2017a). Consider the example of Behaghel et al. (2015), where French firms could opt into an experiment that randomly anonymized the resumes they received from job applicants. The experiment yielded the counterintuitive result that anonymizing resumes *hurt* minority applicants at the selection stage. The authors point to **self-selection into the experiment** as an explanation: their program likely attracted firms that already tend to treat candidates who belong to minorities better, and anonymization prevented these selected firms from treating minority candidates more favorably during the experiment.

[21] This condition is similar in spirit to the Conditional Independence Assumption (Rosenbaum and Rubin 1983).

When the independence condition does not hold (as in Behaghel et al. 2015), researchers must explicitly consider selection into the experiment, in order to derive general conclusions. Alternatively, researchers could conduct NFEs that bypass the selection problem by design and thus allow to recover $\mathbb{E}[y_{1i} - y_{0i}]$ without further assumptions (Al-Ubaydli and List 2013). In this sense, contrary to conventional wisdom, field experiments have the potential for *more control*, and not less, than lab experiments. We return to this point in Section 2.

Note that even when researchers manage to recruit a sample that satisfies the Generalizability Independence Condition above (i.e. p_i is not correlated with outcomes), they can still only generalize to $p_i = 0$ for the subpopulation they draw subjects from, but not necessarily to **other populations** (Deaton and Cartwright 2018). For example, if researchers managed to collaborate with an NGO that has access to a large and representative sample of the population in California (so that the GIC holds), they might able to generalize to those with $p_i = 0$ in California, but not necessarily to the population of Massachusetts or France. This leads us to formulate our fourth threat to generalizability: differences in the populations.

Threat IV: Different populations. Besides characteristics of the experiment (Threat I.), we also need to consider how **characteristics of the population** from which our participants are drawn may affect the generalizability of our results. Even behavior in a stylized and simple game such as the Ultimatum Game exhibits substantial heterogeneity across populations, as seen in a series of AFE conducted in small-scale societies across the world (Henrich et al. 2001). Researchers thus need to discuss how a population different from their experimental sample would react to the same treatment (Athey and Imbens 2017a).[22]

[22] A related dimension to consider is *heterogeneity in response to treatment* across subjects in the study or in the population from which the sample was drawn. We discuss ways to address heterogeneous treatment effect in later sections, through blocking (Section 8) and within-subject design (Section 9).

First, note that if the subject population was a random sample of the "population of interest", then the estimates of the Average Treatment Effect obtained in the experiment generalize to the entire population.[23] Instead researchers often rely on "convenience samples" that are easily accessible to the research team, but the estimates they provide do not necessarily generalize to the entire population (Duflo et al. 2007; Deaton and Cartwright 2018). This problem has been traditionally exacerbated in lab experiments, where subjects are typically from so-called W.E.I.R.D. populations (Western, Educated, Industrialized, Rich and Democratic, Henrich et al. 2010a; Henrich and Heine 2010). The problem of non-representative populations is pervasive in science and not confined to economics: subjects in randomized clinical trials for new drugs are not necessarily a random sample of the population of interest, but are often healthier individuals than the population who is intended to use the drug.[24]

One especially important dimension of generalizability across populations is **gender**: either across men and women, or from one gender to the entire population. Recent years have established a rich and robust literature documenting gender differences in response to a variety of incentive schemes, most notably along the dimensions of competition and risk (Croson and Gneezy 2009), supporting the claim that conclusions drawn from the behavior of members of one gender are unlikely to generalize to the other.[25] The issue of gender becomes even more complex as we take into account its interaction with other covariates.

[23] We loosely define the population of interest as the population for whom we want to obtain the treatment effect estimate; for example, the population targeted by a specific policy.

[24] Travers et al. (For example, 2007, found that less than 10% of asthma patients surveyed qualified for a clinical trial of an asthma medication). For an interesting discussion of heterogeneity in clinical trials, we recommend listening to (or reading the transcript of) the episode "Bad Medicine, Part 2" of the Freakonomics podcast.

[25] Another very stark example concerns the case of clinical trials in the US. In the late 1950s and early 1960s, a drug called thalidomide caused birth defects in hundreds of newborns in a number of countries (Lenz 1988). Although thalidomide was mostly avoided in the US thanks to Frances Oldham Kelsey at the Food and Drug Administration (FDA Bren 2001), more stringent regulations were passed that summarily excluded women from participation in clinical trials (Food and Drug Administration 1997, 2017). Partly as a consequence of those regulations, 8 out of 10 drugs pulled from the market by the FDA in the years 1997-2000 had worse adverse effects for women (Heinrich 2001).

For example, there is evidence that women's preferences over competition change with age such that the gender gap in competition, while large among young adults, disappears in older populations (Flory et al. 2018).

In sum, we urge researchers to carefully consider the limits to the generalizability of their results, and to design their experiments in ways that tackle these four threats to the greatest extent possible. Nevertheless, while generalizability is important to understand and model, we caution against a needless self-destructive overreaction to the generalizability problem that may hinder scientific pursuits. Taken to the extreme, no empirical exercise is perfectly generalizable, so the perfect should not be the enemy of the good.[26] Keeping this balance in mind, in the next section we apply our framework to natural field experiments, and show how they can mitigate or eliminate many potential threats to generalizability.

2 Do more field experiments, especially natural field experiments

This section builds on the framework developed in Section 1 to discuss the advantages and disadvantages of field experiments from the point of view of the generalizability of their results. We first argue that natural field experiments, and to a lesser extent framed field experiments, are often less subject to the threats to generalizability than other types of experiments. We then discuss two typically raised objections to conducting field experiments: lack of control and higher cost, and argue that many times such arguments are confused.

The first threat to generalizability we identified in Section 1 is the change in subjects'

[26] Journals constantly rejecting excellent empirical work on the basis of external validity concerns soon devolves to a *reductio ad absurdum*.

behavior by virtue of being in an experiment and feeling scrutinized: Hawthorne, John Henry, and experimenter demand effects are all commonly used terms describing such potential impacts. In case of overt experiments such as lab, artefactual field (AFE) and framed field experiments (FFE), it is often impossible to rule out these potential biases.[27] On the other hand, the covertness of natural field experiments ensures by design that the environment is natural and there is no sense of scrutiny beyond what is natural in the market of interest, ruling out confounds resulting from a sense of being observed (Al-Ubaydli and List 2013). As a result, there are fewer threats to generalization from direct correlation between y_i and ω in NFEs. In this sense, NFEs are well suited to studying potentially sensitive subjects, such as labor market discrimination (Al-Ubaydli and List 2019).

Researchers conducting FFEs can potentially attenuate the threats to generalizability that result from scrutiny by collecting data over a longer time period, a possibility we discuss in Section 11. Moreover, certain randomized controlled trials, including FFE, can potentially be carried out as single-blind studies where subjects might be aware of being part of an experiment but not of the particular treatments: this is the case when subjects in the control group are given a placebo treatment which they cannot distinguish from the actual treatment (Senn 2013). However, the fact that most economic experiments are not *double-blinded* may introduce biases through the behavior of the researchers who perform the data collection and statistical analysis: even when the participants themselves are not aware of being treated, members of the research team are typically informed of subjects' treatment assignment (Deaton and Cartwright 2018).

The second threat to generalizability is selective noncompliance, i.e. when the probability of changing to another treatment group is different for those who were initially assigned to control versus those who were initially assigned to treatment (for example, in

[27] It may be, however, possible to measure them: see Quidt et al. (2018) for a methodological approach to bounding the experimenter demand effects.

an experiment in which the treatment group is assigned to exercise at the gym, members of the control group might decide to also exercise at the gym). This challenge does not usually apply to lab experiments and AFEs, where noncompliance with one's treatment assignment is typically only possible through leaving the experiment entirely (DellaVigna et al. 2012). Similarly, in NFE subjects are unaware of being assigned to a certain treatment and are thus unlikely to actively try to change their assignment. Switching to a different condition or opting out is often impossible by design (think of Lyft consumers who are randomized into a high or low price for a ride – they receive that price and decide whether to purchase, which is the outcome variable of interest). By their very nature, selective noncompliance is most likely to present problems in certain FFEs.[28]

The third threat to generalizability, that of non-random selection into the experiment, is an aspect where NFEs gain significant attractiveness. By virtue of bypassing the experimental participation decision of subjects altogether, there is no selection by individuals into natural field experiments by design (Al-Ubaydli and List 2013).[29] In lab experiments, it might still be possible to avoid non-random selection into the experiment. For example, Borghans et al. (2009) initially sought volunteers (i.e. those with $p_i = 1$) for their experiment among high-school students, although the experiment was actually compulsory (and included the volunteers). This process avoids non-random selection and also allows for measurement of participation bias, since the researchers know whether p_i is 0 or 1 for

[28]For a mechanism design approach to solving this issue in FFE, see Chassang et al. (2012).

[29]See List (2008) on the ethical considerations behind informed consent. The discussion revolves around benefits and costs, recognizing that for certain sensitive research questions, the subjects' awareness of being part of an experiment may undermine the validity of the research (e.g. measuring the nature and extent of gender or race based discrimination). As List (2008) writes: "This does not suggest that moral principles should be altogether abandoned in the pursuit of science. Quite the opposite: the researcher must weigh whether the research will inflict harm, gauge the extent to which the research benefits others, and determine whether experimental subjects chose the experimental environment of their own volition and are treated justly in the experiment. Local Research Ethics Committees and Institutional Review Boards in the United States serve an important role in monitoring such activities". We would like to emphasize that research can (and should) make participants better off and benefit society, while preserving anonymity and not posing a risk to subject's well-being.

all subjects in the population. However, in practice it is often inconvenient or impossible to make participation in a lab experiments compulsory.

In sum, NFEs are less prone to biases stemming from non-random selection into the experiment, including randomization bias (when subjects are averse to the act of randomization itself), as well as systematic differences in the outcomes or compliance of those who select into the experiment. However, there is an important caveat: even when subjects are unaware of the experiment, there can be participation bias if the participation decision is made on their behalf. For example, if firms selecting to participate in an experiment are such that their employees share a certain characteristic that correlates with the outcome of interest (as in Behaghel et al. 2015), the results from the experiment will only apply to employees of other similar firms. This is because the Generalizability Independence Condition (GIC) derived in Section 1 is violated. In cases where the participation decision is made on behalf of the subjects by another agent, the researchers need to carefully consider whether the GIC holds. If it does not, then statistical intepretation should be adjusted accordingly. This may be the case when the researchers need to collaborate with a number of small self-selected firms, but it can be potentially alleviated when partnering with administrations or large firms who have access to a representative pool of subjects.

The last threat to generalizability applies when we try to extrapolate the findings of one study to a different population. Note that, in this regard, all field experiments (AFE, FFE and NFE) offer an advantage over traditional lab experiments, because they select the population S of interest by design, which is usually different from traditional "W.E.I.R.D. university students" (Henrich et al. 2010b, see also the discussion of Threat IV in Section 1), such as farmers, traders, entrepreneurs, CEOs, physicians, etc. Absent participant selection, within field experiments, NFEs do not have an *inherent* advantage over AFEs and FFEs, in the sense that the population selected S for an NFE can still be very different from the population of interest S', as would happen in AFEs and FFEs. However, NFEs

offer a *potential* advantage because, by collaborating with large entities, researchers can reach a large and often representative sample of the population or the direct population of interest. As an example, consider Hallsworth et al. (2017), who collaborated with a public administration to conduct their tax debtor experiment. Because the population of interest S' over which results should generalize is often the entire population (say, of a given country or region), running NFEs through these types of collaborations allows researchers access to a representative sample. As a result, generalization is either unnecessary (because $S = S'$), or it is feasible either because the subset of interest is part of the experimental population ($S' \subset S$) or because the treatment effect for S' can be extrapolated from a subset of S.

Al-Ubaydli and List (2013) propose a simple framework for generalizability, building on the "all causes model" of Heckman (2000), of which we include the mathematical details in the Appendix, and describe here the main intuition.[30] There are three potential cases of generalizability: zero, local, and global generalizability. Under zero generalizability, results cannot be generalized to any setting different from the one in which they were obtained, which is the most conservative approach. Under local generalizability, results can only be generalized to situations that are very similar to the ones studied in the experiment. Al-Ubaydli and List (2013) argue that under the conservative conditions of zero or local generalizability, field experiments (especially NFEs) can actually offer greater generalizability than lab experiments (and AFEs), because their results can be applied in *some* natural setting (the one in which the experiment was originally performed), for populations and in contexts which would be similar to those of the original experiment. This is especially true if the experiment is implementing a program, and the researchers are evaluating the effects of the program in a particular population. Under global generalizability,

[30] In our model in the Appendix, we use different definitions than Al-Ubaydli and List (2013), but maintain the spirit of the original framework.

on the other hand, results can be extrapolated to contexts that are not necessarily similar to those in which the experiment took place. In this case, neither lab experiments nor field experiments are superior to each other in that they each measure the parameter in the exact situation studied.

One of the most often cited argument against field experiments is the claim that the lab provides more **control** than the field.[31] We agree that lab experiments can have better control over the task subjects agree to participate in, and allow researchers to use induced values (which NFEs by definition have more difficultly doing). However, this alleged disadvantage must be qualified, depending on how we define "field" and "control".

We follow Harrison and List (2004), who view the concept "field" as a continuum, where FFE and NFE are clearly inside the set of field experiments, lab experiments are clearly outside the set, and AFE are somewhere in between. By control, we mean the ability of the researcher to exclude alternative explanations for the outcome, other than the cause of interest. With this definition, the different types of experiments allow for **different types of control**.[32]

NFE could offer more control than lab experiments, not less, along certain important dimensions, the main one being selection into the experiment (Al-Ubaydli and List 2015). As discussed in Section 1, lab experiments, AFE and FFE estimate treatment effects only for those who decide to participate in the experiment ($p_i = 1$), and not for the individuals who do not participate ($p_i = 0$), potentially generating an important bias. Therefore, while the lab provides researchers with *more* control in the environment which participants opt

[31] For example, according to Falk and Heckman (2009) "the laboratory allows tight control of decision environments", while Camerer (2015) claims that "there is little doubt that the quality of control is potentially very high in lab experiments". In a similar vein, Deaton and Cartwright (2018) write: "Exactly what randomization does is frequently lost in the practical literature, and there is often a confusion between perfect control, on the one hand – as in a laboratory experiment or perfect matching with no unobservable causes – and control in expectation – which is what RCTs do".

[32] We elaborate on this point further in Section 3, where we discuss the pros and cons of each type of experiment and the complementarities between them.

into, it provides the researcher with *less* control than NFE over the participation decision (Al-Ubaydli and List 2015). Moreover, while lab experiments are well suited to produce qualitative treatment effects or comparative statics (Levitt and List 2007), under participation bias even qualitative treatment effects are not robust (Slonim et al. 2013).[33] Therefore, when considering the entire experimental situation – from start to finish – NFE could potentially offer more control than lab experiments, because by bypassing the participation decision, they are not subject to participation bias (Al-Ubaydli and List 2013, 2015, see also Section 1).

Despite the several benefits of running FFE and NFE discussed in the paragraphs above, there remains a large obstacle to running more field experiments related to **cost considerations**. As compared to lab experiments, field experiments can be more expensive both in monetary terms and with respect to the planning they require and the time they take to yield results and, ultimately, publications. However, partnering with administrations, NGOs, or firms can substantially reduce the costs of field experiments, and thus result in a win-win collaboration (Levitt and List 2009). Indeed, there are cases when NFE are very low cost, and entail simply the researcher's time when the implementing organization is searching for partners to help generate ideas, design and conduct the experiment.[34] In the limit, it is possible for NFE to incur a negative cost: organizations can realize that the opportunity cost of not knowing the necessary information is too great, and they can actually employ researchers to conduct field experiments that can turn into science (indeed, there is a recent trend at tech companies of hiring PhD economists Athey and Luca 2019).

In summary, field experiments, and especially NFE, offer several advantages over other types of experiments: being covert, they are free of potential bias stemming from experimenter demand effects; they allow for a more complex and natural environment in which

[33] In the sense that the direction of the estimated effect might be opposite to the direction of the true treatment effect.

[34] For a practical take on running field experiments, see (List 2011).

the researcher does not need to know *a priori* all the variables that affect the outcome; subjects belong to the population of interest instead of being W.E.I.R.D. (Henrich et al. 2010a) and, in case of NFE, there is no participation bias because subjects do not self-select into the experiment. All of these features enhance the generalizability of field experiments.

When a researcher decides which type of experiment to conduct (lab, AFE, FFE, NFE), there is a trade-off between the benefits obtained from conducting the experiment (the private benefits to the experimenter, in terms of publication and advancement of her career, and the societal benefit from advancing knowledge) and the cost of running the experiment. In the following section, we discuss this trade-off in more detail in the context of choosing which type of experiment to run.

3 Use lab and field experiments as complementary approaches

After reviewing potential threats to the generalizability of experimental results in Section 1, and discussing what we view as the advantages of field experiments in Section 2, we now tackle the broader question of choosing the right type of experiment (lab experiments, AFE, FFE, or NFE; see the Preliminaries for definitions) for a given research question. Ultimately, we believe that lab and field experiments serve different purposes, and as such they offer **complementarities** in the production of knowledge (Harrison and List 2004; Falk and Heckman 2009). We identify five main issues researchers should consider when choosing between different types of experiments.

First, researchers need to consider the properties of the different types of experiments from the point of view of proper statistical inference (more on this in Section 4). Lab experiments, AFEs, and FFEs can offer more **control on the task** that subjects perform,

once they agree to be in the experiment, than natural field experiments (Al-Ubaydli and List 2013, 2015). This control comes in two forms: i) a more precise environment to establish causation (as an example, consider studies using induced values to test whether prices and quantities converge to neoclassical expectations, as in Smith's (1962) double oral auction lab experiments or List's (2004a) multi-lateral bargaining framed field experiments) and ii) more precise estimates (i.e. lower variance), because one can collect a more homogeneous sample and there are fewer unobservables affecting behavior in the lab, so it is easier to run well-powered studies (see Section 4).

It is also crucial that researchers consider the properties of **replicability**. For example, it has been argued that an advantage of lab experiments is their better replicability (Camerer 2015). Lab experiments can offer a more portable protocol than field experiments, and experimental conditions might be kept constant with more precision. We direct the reader to Section 5 for an extended discussion on the properties of replication.

Combining different types of experiments allows researchers to tackle the issue of **generalizability** by exploring how different factors such as context, scrutiny, stakes and population affect the outcome.[35] As a rule of thumb, the lab is a good place for experiments where the identity of the population does not matter. Gächter (2010) argues that lab experiments using students are excellent as a first step to test economic theories, precisely because most theories assume generality. Neuroeconomic experiments studying brain areas that can be extrapolated to the entire population fit in this category, as well as experiments for which the outcome of interest has been shown to generalize (Stoop et al. 2012; Cleave et al. 2013). AFE, FFE and NFE offer the possibility of using a population of interest instead of a W.E.I.R.D. population (Henrich et al. 2010a,b, also Section 1). FFE and NFE offer the additional benefit of having a natural context, where not only the pop-

[35] Falk and Heckman (2009), discussing the generalizability of experiments, argue that the issue is not necessarily lab vs. field, but "the prevailing conditions such as the details of agent interactions" (see also Section 1).

ulation but also the environment resemble the object of interest. As discussed in the previous section, NFE offer the additional advantage that they bypass subjects' decision of participating in the experiment, therefore avoiding participation bias. This aspect can be especially important if researchers want to scale up their proposed program (Section 12).

Researchers also need to consider the **costs** of running each type of experiment, including all the monetary and logistical costs (recruiting participants and paying them fees, providing treatments and incentives, collecting data, etc.) as well as the opportunity cost of doing other types of research. As we discussed in Section 2, lab experiments are typically (but not always) cheaper than field experiments. Consequently, researchers can often begin by exploring questions using lower-cost lab experiments, and later move into the field to replicate their initial results in a more diverse environment and population. However, this rule of thumb has exceptions. As discussed in Section 2, FFE and NFE can be cheaper (sometimes virtually costless in monetary terms for the researchers) when researchers partner up with governmental agencies, firms and NGOs, creating win-win partnerships (Levitt and List 2009). Moreover, the unit cost per subject can be reduced in field experiments due to economies of scale, and this is compounded with the cost reduction of running experiments in countries with lower costs.

Finally, there are many questions that researchers might simply not be able to tackle in the field, due to ethical or cost **constraints**. To illustrate this point, consider the case of discrimination (Al-Ubaydli and List 2013), where the two main theories in economics are preference-based discrimination (Becker 2010) and statistical discrimination (Arrow 1973; Phelps 1972). Natural field experiments are clearly effective at differentiating between the two potential sources of discrimination, as they target the population and context of interest, and avoid participation bias and experimenter demand effects (for a survey, see List (2006a)). However, the lab can offer a complementary approach to exploring this question: for example, Niederle and Vesterlund (2007) used lab experiments to investigate whether

affirmative action policies affect selection into a tournament, an intervention that would have been difficult to carry out in a natural setting.

In conclusion, different types of experiments offer complementarities in the level of control, replicability and generalizability they allow given their cost, and these trade-offs ultimately determine, for any particular research question, the type of experiment that offers the most value.

4 For proper inference, go beyond p-values

Throughout the previous sections, we focused on the generalizability of experimental results, discussing the extent to which we can extrapolate findings from a given study to other contexts. We now take a step back, and examine how priors should change in light of empirical findings. What conclusions can we draw upon observing a statistically significant result? More generally, what should we consider to be standards of evidence, and what is the framework of proper inference given our research data? We suggest a framework where the benefits from running experiments can be measured by their *informativeness*, i.e. how much they change the priors of the scientific community.[36]

In biomedical and social sciences, including experimental economics, researchers typically obtain their conclusions regarding the existence of an effect or association in their data by conducting (null hypothesis) **significance testing** (Fisher 1925). In particular, they formulate a statistical model complete with a set of assumptions, among them their null hypothesis (H_0, often postulating the absence of the effect/association in question), calculate a test statistic summarizing their data, then compare this statistic to the distribution expected under the model they specified (i.e. assuming that all the model's assumptions, including the null hypothesis, are true). The outcome of this comparison is

[36]This aspect should be considered even when members of the scientific community have multiple priors (see the discussion on priors in this section, and also in Section 8).

summarized in the p-value: the probability that under the specified model the test statistic would be equal to or more extreme than its observed value (Wasserstein and Lazar 2016). A result is then pronounced **statistically significant** if the p-value falls below a pre-specified cut-off (often 0.05, but see the plea from Benjamin et al. (2017) for 0.005). This interpretation, however, is a departure from Fisher's original framework. In his view, significance testing essentially measures the strength of evidence against a null hypothesis, and he leaves the interpretation of the p-value to the researcher. Instead of a strict decision rule, he advocates for examining whether or not the observed p-value is "open to suspicion" – and if so, to run another experiment (Lehmann 1993).

A conceptually different approach to statistical inference is **hypothesis testing**, developed by Neyman and Pearson (1933) with the aim to reduce the subjectivity inherent to Fisher's method. This framework simultaneously addresses the probabilities of two different types of errors of inference: incorrect rejection of a true null (Type I error) and incorrect acceptance of a false null (Type II error). The method requires researchers to formulate a precise alternative hypothesis against which the null hypothesis is tested (in practice, this often means pre-specifying a particular effect size), and to fix in advance the rates of Type I and Type II errors (typically denoted by α and β, respectively). Central to this approach is the concept of **statistical power**: the pre-study probability that the test will correctly reject a false null as a function of the alternative hypothesis (calculated as $1 - \beta$, i.e. 1 minus the Type II error rate). Given the *a priori* specified decision rule (α, β, and the alternative hypothesis H_a), the analysis results in the acceptance or rejection of the null hypothesis. The framework allows for *ex ante* sample size calculations, whereby the researchers assess the number of observations required to detect an effect of the size as stated in the alternative hypothesis, with the pre-specified Type I and II error rates. It is important to point out that hypothesis testing is a *frequentist* approach: it limits the number of mistakes made over several different experiments, but it does not attach an in-

terpretation to a p-value resulting from a single study (Sterne and Smith 2001).

In practice, researchers all too often focus exclusively on the statistical significance of the results when interpreting their findings. Such narrow focus on p-values is dangerous, as it gives rise to several misconceptions. It is crucial to understand that the p-value indicates the incompatibility of the data generated in the experiment with the proposed model, but it does not measure the probability that the null hypothesis is true: recall, the p-value is calculated *under the assumption* that the model is true (Greenland et al. 2016). Thus a p-value of 0.05 from a single study does not ensure that the finding has a mere 5% chance of being a "false positive" (more on false discovery rates later). Furthermore, low p-values should be interpreted as providing evidence against the proposed model *as a whole*, not necessarily against the null hypothesis in particular. Data and model could be incompatible if any of the underlying assumptions are violated, including those related to the quality of measurement, the conduct of the analysis, the reporting of results, etc. Thus, a p-value of a comparison cannot be interpreted in isolation, without considering researcher degrees of freedom and the resulting potential bias (Wasserstein and Lazar 2016). Finally, p-values do not convey any information about the size or importance of the effect in question: tiny effects can produce low p-values if the sample size is large or the precision of the estimate is high enough, and vice versa (Greenland et al. 2016).

Despite repeated calls for moving beyond p-values and examining the statistical power function (McCloskey 1985), most published studies continue to ignore the issue entirely. Ziliak and McCloskey (2004) report that among empirical papers published in the *American Economic Review* in the 1990s, only 8% considered the power of the tests used. More recently, Zhang and Ortmann (2013) failed to find a single study discussing optimal sample size in relation to statistical power among all the articles published in *Experimental Economics* between 2010 and 2012. Given this lack of attention, it is unsurprising that most published studies have very low statistical power. Despite the convention of defining

adequate power as 80%, studies in most fields fall dramatically short of this level.[37] In a survey of more than 6700 studies in empirical economics, Ioannidis et al. (2017) find that the median statistical power of the reviewed research areas is a mere 18%, and nearly 90% of results are under-powered in half of the areas assessed. Coville and Vivalt (2017), focusing on studies in the field of development economics, estimate a median power to detect an average predicted effect of 59%. Only a third of the studies included in their analysis have power greater than 80%. Analyzing time trends, Smaldino and McElreath (2016) find little reason for optimism: according to their survey of review papers published between 1960 and 2011, mean statistical power was 24% in social and behavioral sciences, and showed no increase over time.[38]

Inference from low-powered studies is problematic for at least three reasons. First, by definition, adequate power is required to ensure that studies have a high likelihood of detecting a genuine effect. Low power implies high rates of **false negatives** whereby the null hypothesis of "no effect" is not rejected, despite being false. This aspect is highlighted by De Long and Lang (1992), who review papers published in the 1980s in major economic journals (*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economics and Statistics*) that failed to reject the null hypothesis at the 0.1 level. The authors estimate that in their sample "failures to reject nulls are [...] almost always due to lack of power in the test, and not to the truth of the null hypothesis tested" (De Long and Lang 1992, p.1261). More recently, Coville and Vivalt (2017) estimate an average false negative reporting probability in development economics of approximately 0.53, calculated as the share of incorrectly accepted null hy-

[37] To put it differently, the Type II error rate should be no more than four times the usually prescribed Type I error rate – a convention that is arguably arbitrary and yet routinely followed across different fields of science (Ioannidis et al. 2017). An alternative approach is to simultaneously determine the optimal pair of Type I and II errors according to the circumstances and aim of the specific study, as originally suggested by Neyman and Pearson (1933) and recently reiterated by Ioannidis et al. (2013).

[38] The problem of insufficient power is by no means specific to economics: Button et al. (2013) estimate that the median statistical power in neuroscience is 21%.

potheses over all accepted null hypotheses. Fiedler et al. (2012) argues that researchers' relatively high tolerance for false negatives has potentially irreversible effects on the development of scientific knowledge: since false negative results are less likely to be followed up than false positives, self-correction is less likely to occur in these cases.

A second channel through which low power threatens the credibility of research findings is **effect inflation**: the phenomenon of obtaining "an exaggerated estimate of the magnitude of the effect when a true effect is discovered" (Button et al. 2013, p. 366). This problem is also known as the winner's curse, the Type M error (Gelman and Carlin 2014) or the statistical significance filter (Loken and Gelman 2017). Intuitively, effect inflation occurs because in settings where standard errors are large, only those findings that by chance overestimate the magnitude of the effect will appear statistically significant and thus pass the threshold for discovery. Effect inflation is therefore more severe in underpowered studies that are based on small samples in the presence of high measurement error: studies with power below 50% are likely to yield exaggerated estimates of magnitudes (Gelman and Carlin 2014). In line with this prediction, Ioannidis et al. (2017) estimate that over one-third of the average results of economics research are exaggerated by a factor of more than four, and the majority of reported research is at least twice too large.

The third, less appreciated aspect of statistical power is its relation to **false discoveries**. The connection becomes clear once we abandon the practice of treating a single finding that has achieved formal statistical significance as conclusive evidence, and instead consider a **Bayesian framework of statistical inference** whereby any individual study contributes to scientific knowledge insofar as it moves our priors regarding the existence of the effect/association in question. In this framework, studies may be assessed on the basis of their positive predictive value: the post-study probability that a research finding that has achieved formal statistical significance is indeed true (Wacholder et al. 2004; Ioanni-

dis 2005).[39] The basic ingredients of this metric are the Type I and II error rates (α and β, respectively), together with π, the fraction of true associations among all associations tested in a given field. We treat this fraction as our prior: the pre-study odds that the association in question is true (we discuss different ways to obtain priors later in this section). The post-study probability (PSP) is then defined as the share of *true* associations which are declared true ($(1-\beta)\pi$) divided by the share of *all* associations which are declared true ($(1-\beta)\pi+\alpha(1-\pi)$). As shown in Equation 3 below (reproduced from Maniadis et al. (2014)), the PSP depends on the power of a study $(1-\beta)$ in the following way:

$$\text{PSP} = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)} \tag{3}$$

In particular, since the derivative of Equation 3 with respect to $(1-\beta)$ is positive, the positive predictive value of a study is increasing in its power. As an example, consider a field where π is 0.1 (i.e. 1 out of 10 examined associations is true) and α is fixed at 0.05. Using Equation 3, we find that the post-study probability that a statistically significant finding is genuinely true is 64% in case the level of power is 80%, but falls to a mere 31% for a study with 20% power. Statistically significant results from low-powered studies thus contribute little to scientific knowledge as they lead to a low post-study probability of the findings being genuinely true. Ignoring the above-described framework of inference and treating statistically significant results from underpowered studies as conclusive evidence for the existence of an effect increases the rate of false discoveries, leading to low reproducibility of published results and undermining the credibility of the research field (Munafò et al. 2017; Button et al. 2013). A Bayesian framework also shows that non-significant results, especially when they are obtained in large datasets, can be more *infor-*

[39]The positive predictive value can be understood as the complementary probability of the *false positive reporting probability*, defined by Wacholder et al. (2004) as the probability of no true association given a statistically significant finding.

mative than significant ones. In particular, rejection of a point null hypothesis is often less likely to substantially change priors over a large range of values than is a failure to reject the null (Abadie 2018).

As a side note, we would like to draw attention to the other crucial, yet often overlooked, ingredient of Equation 3: π, the prior probability (or pre-study odds) that the effect being tested exists. The post-study probability that a statistically significant finding is actually true is an increasing function of this prior. Continuing our example above, if the prior we consider changes from 0.1 to 0.01, the PSP falls from 64% to 14% even if the error rates remain fixed at levels conventionally deemed adequate ($\alpha = 0.05$ and $\beta = 0.2$). Consequently, a single "surprise discovery," i.e., the first study to find a statistically significant association in a question where the prior probability was quite low, should only have a limited impact on our post-study belief that the effect actually exists. Given their importance, it is crucial to improve our understanding of priors, and to consider the range of pre-study odds for the question in consideration before running an experiment (Ioannidis 2005). As a general rule, priors are higher in disciplines where empirical research has sound theoretical foundations than in fields where exploratory research is the norm (Maniadis et al. 2017).[40] Abadie (2018) provides a numerical example for constructing a prior distribution for experimental economics studies, using estimates from a replication project we discuss in Section 5 (Camerer et al. 2016; Andrews and Kasy 2017). Obtaining prior probabilities for any particular research question is less than straightforward. One solution is to calculate the PSP using a range of possible values for priors, as demonstrated in e.g. Maniadis et al. (2014). Alternatively, estimates for the pre-study odds may be obtained by consulting experts. As examples, consider Groh et al. (2016) who undertake an "audience expectation elicitation exercise" by collecting treatment effect estimates from members

[40] Card et al. (2011) estimate that 68% of economic field experiments are purely descriptive in the sense that they do not contain even a single line of formal mathematical modeling.

of their audience prior to presenting their results, Coville and Vivalt (2017) who survey a panel of researchers to collect anticipated effects in various development economics studies, or DellaVigna and Pope (2018) who compare expert and non-expert forecasts. Finally, Dreber et al. (2015) use prediction markets to obtain estimates for prior probabilities of specific hypotheses being true.

In the above discussion of the mechanics of statistical inference we have ignored any **"researcher degrees of freedom"** in the design, analysis and reporting that may lead to identifying effects even in the absence of a true association. Recent studies indicate that both specification searching (the practice of trying out several specifications and selectively reporting outcomes that support the researcher's intended conclusion, see e.g. Simmons et al. (2011); Brodeur et al. (2016)) and publication bias (the greater tendency of researchers to submit and editors to publish studies with significant rather than non-significant findings, also known as the "file drawer problem", see e.g. Doucouliagos and Stanley (2013); Christensen and Miguel (2018); Andrews and Kasy (2017)) are prevalent in empirical economics. As Ioannidis (2005) points out, such bias also reduces the post-study probability of a positive finding actually being true.[41] Repeated independent testing by different teams of investigators further lowers the PSP: intuitively, the positive predictive value in this case reflects the fact that only 1 out of n independent studies found a positive association (in Section 5 we discuss how the PSP changes when r out of n independent studies find evidence for the existence of an effect).

In sum, an exclusive reliance on formal statistical significance and inadequate attention to the other ingredients determining a study's positive predictive value (PPV: priors, bias, competition, and, crucially, statistical power) compromise researchers' ability to draw correct inferences from data. Moreover, while reporting the PPV helps emphasize the in-

[41] In the presence of such practices, the positive predictive value may be calculated as follows, where u indicates the extent of bias: $\text{PSP} = \frac{(1-\beta)\pi+\beta\pi u}{(1-\beta)\pi+\beta\pi u+[\alpha+(1-\alpha)u](1-\pi)}$. Maniadis et al. (2017) discuss the determinants of u for a given discipline.

formativeness of a study, it revolves around the "existence" of an effect rather than around effect sizes. Many statisticians call for a departure from methods that focus on *testing* towards those that emphasize *estimation*, such as confidence, credibility or prediction intervals (Wasserstein and Lazar 2016; Munafò et al. 2017) or even "hacking intervals" (Coker et al. 2018).[42]

On the other hand, consumers of research might still find the practice of using a p-value threshold as an established standard for evaluating research findings helpful. Given that the scientific community continues to rely on a universally accepted p-value cut-off, a group of scientists now proposes to make this standard more stringent: Benjamin et al. (2017) argue that novel findings should be labeled as "statistically significant" only if they pass a p-value threshold of 0.005 and recommend treating evidence with p-values between 0.005 and 0.05 merely as "suggestive.". Their proposal promises to reduce false positive rates to acceptable levels in most professions. The proposal sparked an intense debate, with critiques calling for removing (Amrhein and Greenland 2018) or abandoning (McShane et al. 2019), rather than redefining, statistical significance, and suggesting a new approach to reporting results whereby researchers transparently present and justify their design choices, including their chosen significance level (Lakens et al. 2018).

This lively debate signals a growing interest among experimental scientists in the issue of proper inference, a welcome development that we hope will translate into actual changes in practices and norms in the scientific community. In the following sections we review several practical recommendations that have the potential to substantially improve the reliability of scientific results. In Section 5 we begin with what we see as the most pressing issue currently: we discuss the importance of replications, and present incentive-compatible methods to encourage them.

[42]Sterck (2018)suggests assessing the economic importance of a regressor by measuring the percentage contribution of the given explanatory variable to the variation in the dependent variable, and offers two alternatives for handling the variation induced by explanatory variables that are correlated.

5 Replicate early and often

We believe that the best approach to increasing the reliability of results from experimental economics lies in replication. Recent controversies surrounding topics such as ego depletion in the psychology literature (Hagger and Chatzisarantis 2016; Hagger et al. 2010; Carter and McCullough 2014) or the impact of de-worming programs in development economics (Croke et al. 2016) all highlight the importance of replications. In the following we define what we consider replication, and demonstrate using a Bayesian framework of inference why it is crucial for the credibility of science. We then discuss what the "natural rate of replication" and the rate of reproducibility in economics are today. Additionally, we review several proposals to incentivize replication.[43]

As Clemens (2015) points out, there is currently no universally accepted standard in economics as to what exactly constitutes a replication. Levitt and List (2009) propose definitions that are well-suited to experimental studies. In the most narrow interpretation, a replication means taking the original data generated by an experiment and re-analyzing it to confirm the original findings. In the terminology of Hamermesh (2007), this would constitute a **pure replication:** examining the same question and model using the underlying original data set. This approach may help to address issues with the internal validity of a study, for instance through uncovering coding errors or mistakes in calculations.[44] A broader interpretation of replication in experiments involves running a new experiment closely following the original protocol to test whether similar results can be generated using a new subject pool. Such a study would be classified as **statistical replication:** based on a different sample, but using an identical model and underlying population

[43] Other valuable methods aimed to serve the goal of "research synthesis" are literature surveys and meta-analyses; for reviews on these methods, refer to e.g. Anderson and Kichkha (2017); Maniadis et al. (2017); Maniadis and Tufano (2017).

[44] Even without an explicit mistake on the researchers' side, empirical results are not necessarily robust; as an example, consider McCullough and Vinod (2003) who report that nonlinear maximization methods from different software packages often produce wildly different estimates.

(Hamermesh 2007). This method has the potential to fix sampling errors or insufficient power. Finally, the third and broadest category entails testing the hypotheses of the original study using a new research design. This characterization is a **scientific replication** according to Hamermesh (2007), as it involves a different sample, a different population, a different situation, and a perhaps similar but not identical model. These replications help assess the robustness of the original finding, and may inform discussions on the generalizability of the original result (see Section 1).[45]

To illustrate why statistical replications are crucial, let us return to the Bayesian framework of inference introduced in Section 4. Equation 3 presented the post-study probability (PSP) of a finding actually being true, conditional on *a single study* providing statistical evidence in favor of its existence. Following Moonesinghe et al. (2007), we can adapt this formula to calculate the PSP when at least r out of n independent studies find a significant result for the association in question. As before, we obtain the PSP as the fraction of true associations declared true over all associations declared true:

$$\text{PSP} = \frac{\pi \sum_{i=r}^{n} \binom{n}{i} (1-\beta)^i \beta^{(n-i)}}{\pi \sum_{i=r}^{n} \binom{n}{i} (1-\beta)^i \beta^{(n-i)} + (1-\pi) \sum_{i=r}^{n} \binom{n}{i} \alpha^i (1-\alpha)^{(n-i)}} \tag{4}$$

Using Formula 4, Moonesinghe et al. (2007) and Maniadis et al. (2014) demonstrate how a few successful replications can increase the positive predictive value of a finding. This increase is particularly dramatic in cases when prior probabilities are low.[46] Within the same framework, Coffman and Niederle (2015) argue that even the most inaccurate beliefs

[45] Clemens (2015) suggests an alternative classification, differentiating between replication tests (including verification and reproduction tests) and robustness tests (including reanalysis and extension tests). See also the discussion in Duvendack et al. (2017).

[46] Consider the following example, based on Maniadis et al. (2014): $n = 15$ researchers independently run the same study with 80% power to detect an association that has a 10% prior probability of being true. When a single study out of the fifteen attempts finds a statistically significant association (i.e. $r = 1$), then the post-study probability that this positive finding is actually true is a mere 17% (remember that the corresponding PSP was well above 50% in the absence of researcher competition (i.e. $n = 1$), see Section 4). However, the post-study probability that the association in question really exists increases to over 90% in case of just two successful replications.

can be corrected within three to five replications.[47]

Despite a general consensus among economists regarding the importance of replication, it remains largely "an ideal to be professed but not practiced" (Mueller-Langer et al. 2019). Incentives for individual researchers to replicate a project or to have their own work replicated are low or missing entirely. Replications typically bring little recognition for their authors despite the substantial work they entail. The process is particularly tedious because data and code for published articles are often unavailable – even though most leading economics journals have introduced data sharing requirements and mandatory data archives, such policies are not necessarily enforced (Höffler 2017). As Duvendack et al. (2017) observe, replications are usually regarded as unoriginal or "derivative". Worse, they may ignite animosity among researchers if authors of the original work treat replication attempts as threats. Moreover, journals may be reluctant to publish replication studies for fear of not receiving enough citations (Duvendack et al. 2017). Indeed, according to a survey by Mueller-Langer et al. (2019), from 1974 to 2014 less than 0.1% of publications in the top-50 economics journals were replications. Given the difficulties of publishing a 'mere replication,' conducting an extension study where the control treatment replicates a previous finding is often a more attractive alternative. This, however, makes replications hard to identify: as Coffman et al. (2017) point out, such 'implicit replications' are often reported as part of a paper with a much larger scope, without being labeled as replications. Other times, successful replications are simply not considered interesting enough to be published. As a result, it is less than straightforward to assess how often replications actually occur.

A few recent papers attempt to address this issue and estimate the "natural rate of replication" in economics. First, Berry et al. (2017) focus on all the empirical papers pub-

[47]For a more nuanced approach that takes into account various forms of researcher bias among the replicators, see Maniadis et al. (2017).

lished in the centenary volume (2010) of the *American Economic Review*, and manually code all their published citations as either replications, robustness tests, extensions, or none of the above. They find that less than a third of the 70 papers have been replicated at least once, where a replication is defined as a project "speaking directly to the veracity of the original paper." Validating the assertion that the visibility of replications is low, Berry et al. (2017) find considerable uncertainty among the authors of the original papers over the number of extant replications of their studies. Second, Sukhtankar (2017) analyzes 1056 empirical papers in development economics published in the top ten general interest journals between 2000 through 2015, perform a reverse citation search, then search within the ensuing list for "replication" or alternative cognates. His results suggest that only 5.4% of the studies in their sample were replicated in a published paper or a working paper, the rate being higher (12.5%) for studies based on randomized controlled trials. Third, Hamermesh (2017) collects ten leading papers from labor economics with at least 20 years of citation history, and classifies their citing papers as either i) related to, ii) inspired by, iii) very similar to but using different data, or iv) a direct replication at least partly using the same data. He finds that of the more than 3000 citing studies, only 0.6% fall into the last category. On the other hand, 7 out of the 10 original studies he surveyed were replicated at least five times, and all of them at least once. Finally, Maniadis et al. (2017) survey experimental papers published between 1975–2014 in the top 150 journals in economics, and estimate that the fraction of replication studies among all experimental papers in their sample is 4.2% (taking into account 'implicit replications' as well). Overall, these studies suggest that the natural rate of replication in empirical economics is low, although heavily cited and influential papers do tend to get replicated.

The above results concern the rate at which replications are *attempted*, leaving aside the question of what share of these replications is *positive*, i.e. confirm the findings of the original study. Measuring rates of reproducibility in economics dates back to the quest of

Dewald et al. (1986) to replicate findings from articles published in the *Journal of Money, Credit and Banking*. The authors famously concluded that inadvertent errors were a "commonplace rather than a rare occurrence." Another key insight of the Dewald et al. (1986) study was the alarmingly high share of authors who were unwilling or unable to supply their data and code to the replicators. According to Chang and Li (2017), this problem is still pervasive: in their attempt to replicate macroeconomic papers published in 13 well-regarded journals, the greatest obstacle they faced was authors' failure to provide their data and code files. As a result, they were only able to qualitatively reproduce the key results of 29 out of the 59 papers they sampled.

Focusing on experimental economics specifically, Deck et al. (2015) review several replication attempts, mostly in the context of public goods provision, with varying outcomes. The first systematic evidence of replicability of laboratory experiments in economics is provided by Camerer et al. (2016) who replicate 18 studies published in the *American Economic Review* and the *Quarterly Journal of Economics* between 2011 and 2014, according to pre-analysis plans posted prior to conducting the replication studies. They find a significant effect in the same direction as in the original study in 11 out of the 18 studies, corresponding to a reproducibility rate of 61%.[48] They also discuss alternative replication indicators, e.g. whether the 95% confidence interval of the replication effect size includes the original effect size, or whether the replicated effect lies in a 95% prediction interval. These measures suggest higher rates of replicability (66.7% and 83.3%, respectively). The authors also compare the replicated effect sizes with the original, and find a mean relative effect size of 65.9%. The finding that the replicated effect sizes tend to be smaller than the

[48]While lower than desirable, this share is considerably higher than the replicability rates uncovered in the Reproducibility Project: Psychology (RPP) (Open Science Collaboration 2015), a project that involved replicating 100 studies published in three psychology journals. Their results paint a rather grim picture of the reliability of psychological research: while 97% of the original studies found significant results, only 36% of the replications were able to reproduce these significant findings. In the Many Labs 2 project, 15 of the 28 attempted replications provided evidence in the same direction as the original finding and statistically significant at the 5% level (Klein et al. 2018).

original ones reflects the effect size inflation phenomenon discussed in Section 4. Overall, Camerer et al. (2016) interpret their findings as suggesting "relatively good replicability of results."

Another noteworthy replication initiative is the Social Sciences Replication Project, whose collaborators aimed to replicate 21 experimental studies in the social sciences published in the prestigious journals *Nature* and *Science* between 2010 and 2015. They find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size (Camerer et al. 2018). Finally, while both of the above-mentioned projects focus on results published in top journals, Maniadis et al. (2017) analyze replication attempts from 150 economic journals, and find a "success rate" of 42.3% among the 85 experimental replication studies in their sample.

The recent surge of interest in reproducibility also ignited an intense discussion about the most effective ways to incentivize replications. We conclude this section by reviewing a few suggestions that we find particularly promising. The first set of ideas addresses the current difficulty of publishing replication studies, suggesting the creation of a specific outlet in the form of a new journal dedicated to replications (Coffman and Niederle 2015), or including one-page "replication reports" in top journals (Coffman et al. 2017). The recent launch of the *Journal of the Economic Science Association*, with a special section devoted to replications, is a promising step in this direction. These suggestions could be especially effective coupled with a new norm that requires citing replication work alongside the original, increasing the returns both to the publishing journals and to the authors of the replications. In addition, departments should consider systematically incorporating into their hiring and tenure decisions an assessment of whether a researcher has promoted transparency in their career through replication and data sharing.

Second, Maniadis et al. (2015) emphasize the need to change authors' incentives to col-

laborate with replicators.[49] In their view, journals should always allow original authors to give their commentaries after a replication attempt; they also suggest considering the number of replication attempts as a metric for one's research quality. Third, Butera and List (2017) design a new, incentive-compatible mechanism whereby the original investigators of a study commit to only publishing their results as a working paper, and offer coauthorship of a second paper (submitted to a peer-reviewed journal) to other researchers who are willing to independently replicate their experimental protocol in their own research facilities. This mechanism allows the original authors to signal the ownership of the research idea, while ensuring the credibility of their results (in case they indeed replicate). At the same time, scholars on the team of replicators, in return for bearing the cost of replications, would benefit from coauthoring a novel study. Finally, Dreber et al. (2015) suggest using prediction markets with experts as quick and low cost ways to obtain information about reproducibility.[50]

Combined, these approaches have the potential to make replication more prevalent by increasing its attractiveness to researchers. Such a shift in attitudes could also have positive consequences on how research is conducted in the first place: as Duvendack et al. (2017) point out, replication may have a deterrent effect on questionable or fraudulent research practices by increasing the likelihood that such practices will be discovered. The profession as a whole could benefit from a culture that recognizes the intrinsic value of replications.

In sum, replication serves to prevent, expose and correct wrong inferences, and is thus

[49] See also Maniadis et al. (2017) for a systematic review on the problem of information revelation in science.

[50] Dreber et al. (2015) set up prediction markets in conjunction with the Reproducibility Project: Psychology (described in footnote 48) where participants could bet on the success of the attempted replications. Prediction markets were found to predict the outcomes of the replications well, performing better than a survey of participants' individual forecasts. While Camerer et al. (2016) confirm the result that beliefs elicited though a prediction markets are positively correlated with a successful replications, they do not find that this method works better than belief elicitation through a survey.

inevitable for producing empirical results that can reliably form the basis of both economic theory and policy. However, replication does not eliminate the need for well-powered studies: replication projects with low statistical power contribute little to the evidence in favor of or against a hypothesis. In the next section, we follow this line of reasoning by urging researchers to perform ex ante power calculations and to design their experiments in ways that maximize statistical power.

6 Consider statistical power in the design phase

As discussed in Section 4, insufficient statistical power in experiments poses a major challenge to proper inference. The most straightforward remedy, of course, is to avoid conducting low-powered studies in the first place. In case of experiments, this requires taking the question of statistical power seriously in the design phase. In the following, we describe the basic principles of optimal sample size calculations, and then review sample arrangement practices that maximize power given the available budget. The section is intended as a overview of the most important considerations; the interested reader can find more details in List et al. (2011), Duflo et al. (2007) and Cox and Reid (2000).

Power calculations (i.e. the assessment of the precision of inferences expected to be achieved with a given sample size), or optimal sample size calculations (i.e. the estimation of the sample size required to attain a certain precision), are crucial steps *prior* to conducting an experiment (Gelman and Hill 2007).[51] Power calculations are most often advocated as tools for preventing high rates of false negatives. They also increase efficiency by ensuring that scarce resources are not wasted on studies that are larger than necessary.

[51] Gelman and Carlin (2014) go a step further and suggest performing what they call a "design analysis," complementing power calculations with an assessment of the sign error rates (the probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero) and the exaggeration ratio (the expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero).

Moreover, pre-determining sample sizes can curb bias by reducing experimenters' temptation to collect more data when initial results are insignificant but "go in the right direction" – a practice that could lead to high rates of false positives (Zhang and Ortmann 2013). Despite these arguments, in practice researchers often forgo *ex ante* sample size calculations and rely on shortcuts with little theoretical justification when designing their experiments.[52]

As discussed in Section 4, sample size calculations are rooted in the framework of hypothesis testing. As such, they require researchers to specify (1) a null hypothesis and an alternative hypothesis, (2) the desired significance level and power of the test and (3) the statistical test to be used in the subsequent analysis (List et al. 2011). These considerations allow the researcher to simultaneously control the likelihood of committing either a Type I or a Type II error. In particular, by considering the hypothetical distributions of the test statistic under the null and the alternative hypothesis, the researcher obtains critical values for the test statistic corresponding to the pre-specified error rates. The null hypothesis for a test is typically specified as no effect/association, while the alternative hypothesis typically postulates that the effect size is at least as large as a specific value. Alternatively, for a given budget and thus fixed sample size, one can calculate the minimum detectable effect size given the pre-specified acceptable error rates.

While the researcher has discretion over these three building blocks of power calculations (hypotheses; acceptable error rates; the test used for comparison), translating critical values to optimal sample size requires knowledge of the variance of the outcome – a parameter unknown prior to conducting the experiment. This feature makes *ex ante* power calculations inherently hypothetical, as they are based on the researcher's expectations about the underlying data generating process (Gelman and Carlin 2014). However, one

[52]List et al. (2011) mention the practice of assigning 30 subjects to each treatment arm as an example for a widely used yet theoretically unfounded rule-of-thumb.

should not use the hypothetical nature of power calculations as an excuse for skipping this step in the design phase. Researchers can use data from previous experiments or pilot studies to form beliefs about the variance of outcomes. It is also instructive to calculate the statistical power for a range of different hypothesized values of the variance. When deciding what effect size to target, researchers should consider what difference is actually practically or economically relevant – an aspect that is still largely overlooked both at the design and the inference stage.[53] A useful practice is to express minimum detectable effect sizes in terms of standard deviation changes to facilitate comparison with existing studies in the field (e.g., the researcher may desire to have her experiment detect a 0.1 standard deviation treatment effect).

We demonstrate the framework of power calculations through a simple example adapted from Section 3.1 of List et al. (2011). Suppose we are interested in estimating the average treatment effect from an experiment where participants are randomly assigned to either the treatment or the control group. For now, assume that we only test a single hypothesis in our study (more on multiple comparisons later). For simplicity, we assume that our data is generated by the following model:

$$Y_i = \beta + \tau D_i + \epsilon_i,$$

where Y_i is a continuous outcome variable, D_i is a binary treatment indicator, the estimated treatment effect is homogeneous, and ϵ_i is an idiosyncratic error term with variance σ^2_ϵ. Throughout this example, we assume that the unobserved components of outcomes are independently distributed among our subjects, and relegate the discussion of inference with grouped errors to later in the section. Errors may be heteroscedastic: we allow the

[53] Ziliak and McCloskey (2004) review papers published in the American Economic Review and find that the share of papers discussing effect sizes rather than merely the significance (and maybe the sign of the estimated coefficients) is still low.

variances of the error term in the control (σ_C^2) and the treatment conditions (σ_T^2) to vary. Assuming normality, we use a two-sided t-test for comparing the means of the outcome variable between the groups. These assumptions allow us to derive simple and intuitive closed-form solutions for the optimal sample size or the minimum detectable effect size.

There are n_C and n_T subjects in the control and the treatment groups. Due to random assignment, the estimated average treatment effect is obtained simply as the difference in means between the treatment and the control group: $\hat{\tau} = \bar{Y_T} - \bar{Y_C}$, with variance: $\hat{V} = \text{Var}(\bar{Y_T}) + \text{Var}(\bar{Y_C}) - 2\text{Cov}(\text{Var}(\bar{Y_T}), \text{Var}(\bar{Y_C})) = \sigma_T^2/n_T + \sigma_C^2/n_C$.[54] Our goal in this exercise is to determine the smallest true effect size we can detect given our sample size and required statistical significance and power.

To begin, let us assume that the null hypothesis is true: the true average treatment effect is zero. The hypothetical distribution of the estimated treatment effects is then centered around zero, as shown in the top panel of Figure 1. As aforementioned, in order to control the Type I error at a rate of α, we reject the null hypothesis only if we observe a t-statistic that is equal to or more extreme than our critical value. Equation 5 summarizes this condition: the left hand side of the equation is the t statistic estimated from a test comparing the means of the outcome variable in the control and the treatment group assuming that the true average treatment effect is zero, while $t_{\alpha/2}$ is the critical value corresponding to a false positive rate of α in a two-sided test.

$$\frac{\bar{Y_T} - \bar{Y_C}}{\sqrt{\frac{\sigma_C^2}{n_C} + \frac{\sigma_T^2}{n_T}}} \geq t_{\alpha/2} \tag{5}$$

Now consider the distribution of the treatment effect under the alternative hypothesis,

[54] Note that the true variances in the treatment and control groups are typically unknown a priori and are themselves estimated from the data, often by means of the Neyman variance estimator, a conservative randomization-based approach (Samii and Aronow 2012).

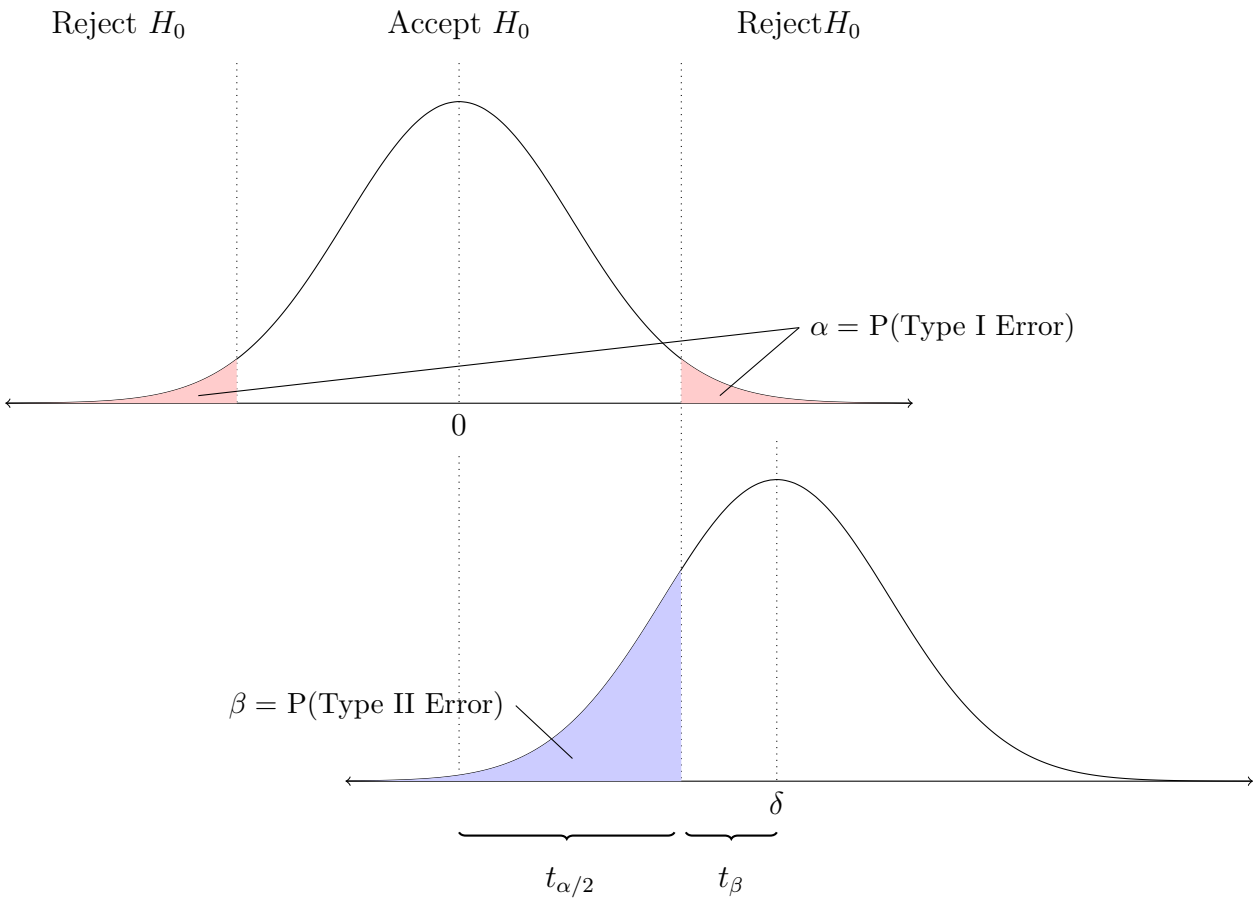


Figure 1: Hypothetical distributions of the estimated treatment effect under H_0 and H_a

assuming a true effect size of δ. The hypothetical distribution of the estimated treatment effects under this alternative hypothesis is shown in the bottom panel of Figure 1. The power of our test to identify a true effect size of δ can be thought of as the fraction of the area under this distribution that falls to the right of the critical value $t_{\alpha/2}$: this is the region where we correctly reject the null hypothesis. Limiting the Type II error rate to β (resulting in a statistical power of $1 - \beta$ for our test), we can calculate the **minimum detectable effect size** of our experiment: the smallest value for which we can (correctly) reject the null hypothesis of no treatment effect with probability $1 - \beta$ at a significance level of α. This minimum detectable effect size δ_{min} can be expressed as a function of the

sample sizes and variances in the control and treatment groups as:

$$\delta_{min} = (t_{\alpha/2} + t_{\beta})\sqrt{\frac{\sigma_C^2}{n_C} + \frac{\sigma_T^2}{n_T}} = (t_{\alpha/2} + t_{\beta})\sqrt{\hat{V}} \tag{6}$$

Equation 6 shows that the lower the variance of the treatment effect estimator, the smaller the effect size we can detect. This estimated variance, in turn, depends on the sample sizes and the variances of the error terms in the two groups.[55] Equation 6 can be re-arranged to determine the sample sizes in the treatment and the control group that are required to detect a treatment effect of the size of δ given the variance of the estimator and the pre-specified Type I and II error rates.

Analytical power calculations such as the example presented above are useful for simple comparisons. For non-parametric tests and more complex or more specific design choices, simulation-based power calculations provide more flexibility. These approaches require the researcher to specify the underlying model complete with the experimental design and sample sizes, the values of the covariates, the parameter values expressing the distribution of the outcome variable under the alternative hypothesis, and the variances (Feiveson 2002). Based on this model, the researchers generate their synthetic data and run their estimation on these data a large number of times, obtaining a p-value in each round of the simulations. Power is then calculated as the proportion of p-values that are lower than the pre-specified cutoff value α. Several recent papers provide more details along with software packages that implement simulation-based power calculations (Feiveson 2002; Luedicke 2013; Bellemare et al. 2016; Burlig et al. 2017).

While often not flexible enough for practical purposes, Equation 6 can still provide

[55] As a numerical example, consider an experiment where a total of 2000 participants are equally divided between a treatment and a control group ($n_C = n_T = 1000$), and we assume equal variances in the two groups ($\sigma_C = \sigma_T = \sigma$). Assuming $\alpha = 0.05$ and $\beta = 0.2$, using a two-sided t-test we can then detect a minimum effect size of 0.125 standard deviation: $\delta_{min} = (t_{0.025} + t_{0.2})\sqrt{\frac{\sigma^2}{1000} + \frac{\sigma^2}{1000}} = (1.96 + 0.84)\sqrt{1/500}\sigma \approx 0.125\sigma$

valuable insights for the derivation of basic heuristics intended to maximize the precision of a study through design. A straightforward way to increase power is to increase the number of observations – this, however, is often impossible due to budget constraints or other practical considerations. We thus focus on sample arrangement techniques for a given experimental budget that aim to reduce the variance of the estimate through other channels.

The first such rule concerns the assignment of subject to treatment or control groups. While it is common to assign equal number of subjects to all conditions, studying Equation 6 we find that this practice is only optimal in case we expect the variances to be the same across groups. Otherwise, the ideal ratio of the sample sizes assigned to treatment and control is equal to the ratio of the standard deviation of outcomes in the two groups. For the special case of a binary outcome variable, the same logic implies that sample sizes in the treatment and control groups should only be equal in case the null hypothesis predicts equal means – and thus equal variances – between the two groups.

The optimal design also takes potential heterogeneity in data collection costs into account to maximize power for a given experimental budget. The unit cost of obtaining an extra subject might differ between treatment and control groups. Intuitively, providing a treatment is often a lot more costly than simply surveying someone or relying on administrative data in the control condition (Duflo et al. 2007). It can be shown that the optimal share of subjects assigned to treatment versus control is inversely proportional to the square root of the respective sampling unit costs (List et al. 2011).

So far we have focused on experiments with a binary treatment indicator. In studies where the experimenter can choose different levels of treatment intensity, precision may be increased through design that maximizes the variance of the treatment variable. In particular, the number of subjects allocated to the different levels of treatment should reflect our priors of the functional form of the treatment effect. Identification requires the num-

ber of treatment cells used to be equal to the highest polynomial order plus one (List et al. 2011). For instance, assuming a quadratic treatment effect, we should allocate subjects to three treatment cells at the two extremes and at the midpoint of the feasible range, assigning one-fourth of subjects to each extreme and half to the midpoint for maximum variation in the treatment variable.

Power calculations also need to account for the randomization technique used to assign subjects to specific treatments. Cluster-randomized designs, where the unit of randomization does not coincide with the unit of analysis, are commonly used in field experiments. For instance, even if student-level data is available, institutional constraints or fear of spillovers might induce researchers to randomize at the classroom or school level. In such cases, our assumption of i.i.d. error terms is often not justified. Clustered designs therefore necessitate power analysis that accounts for group-level shocks (see Duflo et al. (2007) who derive the expression for the variance of the estimated treatment effect in cluster-randomized designs, Abadie et al. (2017) on when to adjust standard errors for clustering, and Young (2016) for a degree-of-freedom correction for robust and clustered covariance matrix estimates). Optimal design in such experiments balances two opposing forces: for a given number of experimental subjects, increasing the number of clusters we sample from leads to greater gains in power than sampling additional individuals from already included clusters. However, adding a participant from a new cluster tends to be more expensive than another participant from an existing cluster. Additionally, Chandar et al. (2018) argue that in case of heterogeneous treatment effects at the cluster level, the researcher may want to include more treated clusters than control clusters.[56] We return to the question of treatment assignment in Sections 8 and 9 where we discuss in detail the practice of blocked randomization and within subject designs.

[56]The intuition is that if the researcher's intervention leads to different effects across different clusters, having more treated clusters can help average over those differences and recover the mean effect (Chandar et al. 2018).

Finally, we would like to draw attention to an implicit assumption we made throughout this section: we assumed that participants comply with their treatment assignment. However, as we have discussed in Section 1, compliance is often imperfect. In overt field experiments that randomize *access* to a particular program or service (common in development economics), take-up of the offered treatment is often low, jeopardizing researchers' ability to detect the impact of the program. McKenzie (2011) points out that the sample size required to detect a given change resulting from a treatment is inversely proportional to the difference in the proportion of the treatment group that takes up a given intervention relative to the control group.

We end this section by re-iterating an important qualification to the above-described framework: it discusses power calculations in cases of a single comparison, whereas most studies test multiple hypotheses. In the following section, we discuss different manifestations of the multiple comparisons problem, and show how the experimental design and the statistical analysis should be modified to account for multiple hypothesis testing.

7 Adjust for multiple hypothesis testing, in power tests and in data analysis

In the previous section we derived our results based on the assumption that researchers evaluate a single hypothesis. In practice, however, most research in applied economics entails more than one comparison performed within the same study. **Multiple hypothesis testing (MHT)**, or the multiple comparisons problem, refers to the practice of simultaneously considering multiple statistical inferences (Miller 1981). Failure to account and correct for multiple hypothesis testing increases the likelihood of false positives and contributes to the replicability crisis of the social sciences (List et al. 2019b). As an example, consider a study wherein a researcher jointly tests N mutually independent hypotheses, all

of which are true and therefore should be accepted. Fixing the Type I error rate for a single comparison at a level α, the probability of at least one false rejection among *all* comparisons in this case is $1-(1-\alpha)^N$. Setting $\alpha = 0.05$, the probability of observing at least one false positive is over 14% in case of just three hypotheses, and it exceeds 50% when testing 14 or more hypotheses. In the following, we provide an overview of the prevalence of the problem in the literature, and discuss possible solutions.

The practice of ignoring multiple hypothesis testing corrections is pervasive in experimental social sciences. List et al. (2019b) differentiate between three main cases of multiple hypothesis testing. The most common occurrence involves analyzing the impact of an intervention on **multiple outcomes**. According to an overview by Anderson (2008), 81% of surveyed papers published from 2004 to 2006 report results on at least five outcomes, and a striking 61% consider ten or more outcomes (the number of *unreported* comparisons is likely to be even higher). Yet only 7% of these papers account for multiple hypothesis testing in their inference. The second widespread form of MHT entails comparisons across **multiple subgroups** of the study population. Analyzing heterogeneous response to treatment by gender, ethnicity, age, etc. falls into this category. Finally, analyzing experiments with **multiple treatments** (either estimating the effect of each treatment condition versus a control, or performing all possible pairwise comparisons across multiple treatments and a control) also constitutes a case of MHT.

There are three main approaches to managing the multiple hypothesis testing problem: (1) reduce the number of comparisons carried out, (2) use machine learning (ML) techniques to deal with several different outcomes or dimensions of heterogeneity in a flexible and principled way, and/or (3) adjust the statistical inference to take into account the family of hypotheses considered in the analysis. The first approach involves restricting the analysis to a specific set of outcomes based on *a priori* notions of importance, and/or using **summary index tests** that pool multiple outcomes into a single measure (Ander-

son 2008). The second approach makes use of machine learning techniques by reframing the problem as a *prediction* rather than an estimation task. Instead of testing whether the treatment affected different outcome measures, ML techniques instead ask whether treatment assignment can be predicted from a set of observable outcomes (Mullainathan and Spiess 2017). Similarly, rather than testing multiple potential dimensions of heterogeneity specified by the researcher, ML techniques seek to identify subgroups such that treatment effects are similar within and different across groups, allowing the researcher to discover more flexible forms of heterogeneity (Athey and Imbens 2017b).[57]

The third method (the focus of this section) accounts for the multitude of tests carried out by adjusting the inference from the analysis. Multiple testing procedures often control the **family-wise error rate**: the probability of rejecting at least one true null hypothesis among a set of hypotheses we jointly test (Heckman et al. 2010).[58] Alternatively, when the number of hypotheses tested is very large, researchers often choose instead to control the m-familywise error rate (the probability of m or more false rejections), the tail probability of the false discovery proportion (the fraction of false rejections), or the false discovery rate (the expectation of the proportion of rejected true null hypotheses among the rejected hypotheses) (Benjamini and Hochberg 1995; Benjamini et al. 2006; List et al. 2019b)

Different techniques have been developed to adjust the standards of inference to take into account multiple hypothesis testing. **Single-step procedures** simultaneously compare all the individual test statistics from the different comparisons to their critical values. Often (though not always) the same critical value is used for all comparisons. As an example, consider the most well known multiple testing procedure developed by Bonferroni

[57]For correct inference, Athey and Imbens (2016) propose an "honest" approach, whereby one sample is used to divide participants into subgroups and another to estimate treatment effects for each subgroup. We further discuss the use of ML techniques for studying heterogeneous treatment effects in Section 9.

[58]What constitutes a "family" of comparisons is not always straightforward to determine. In general, the decision should be guided by the conceptual/theoretical similarity of the tests. For a helpful discussion, we recommend the related blog post by Daniel Lakens: "Why you don't need to adjust your alpha level for all tests you'll do in your lifetime".

(1935), applied to the calculation of confidence intervals by Dunn (1961). This technique consists of computing an individual p-value for each hypothesis tested, and rejecting a hypothesis only if its p-value does not exceed α/S, where S is the total number of comparisons performed. Under the assumption that the null distribution of each p-value is uniform, this method asymptotically controls the family-wise error rate at level α (Romano and Wolf 2005).

Stepwise methods of multiple testing procedures also start with a single-step method. However, instead of stopping after the first set of comparisons, these methods allow the researchers to reject further hypotheses in subsequent steps by decreasing the critical values for the remaining hypotheses, taking into account the hypotheses already rejected in previous steps. The methods continue until no further hypotheses are rejected (Romano and Wolf 2010). Stepwise procedures can be further classified into stepdown and stepup methods. Stepdown methods begin by considering the most significant hypotheses, and then continue to evaluate hypotheses with smaller test statistics. Romano and Wolf (2010) show that the classical method of Holm (1979) can be formulated as a stepdown procedure where the criterion for rejection for the most significant hypothesis is the same as in the Bonferroni-method, but the criteria get less strict for larger p-values.[59]

The appeal of the traditional methods of Bonferroni (1935) and Holm (1979) lie in their simplicity – however, they are often overly conservative. Procedures with more power to reject false null hypotheses have been designed by taking into account the joint dependence structure of the individual p-values (e.g. Romano and Wolf 2005; Heckman et al. 2010).[60] Based on Romano and Wolf (2010), List et al. (2019b) developed a bootstrap multiple hypothesis testing procedure that asymptotically controls the family-wise error rate under fairly weak assumptions. Their procedure was designed to simultaneously han-

[59]For an example for a stepup procedure, refer to e.g. Benjamini and Hochberg (1995).

[60]For a discussion on the meaning of power in a multiple hypothesis testing context, refer to Romano and Wolf (2005).

dle all three scenarios of MHT in experimental economics discussed above. The method is asymptotically balanced in that the marginal probability of rejecting any true null hypothesis is approximately equal in large samples, and provides an improvement over classical methods in terms of power through incorporating information on the dependence structure (and can lead to further gains by exploiting logical restrictions across null hypotheses in case of multiple treatment arms).[61]

Accounting for multiple hypothesis testing often leads to different conclusions than inference that ignores the multitude of tests carried out at once. For instance, Lee and Shaikh (2014) demonstrate how the significance of PROGRESA's estimated effects on school enrollment across sub-populations change once we account for multiple inferences. While demonstrating their approach to correcting for MHT, List et al. (2019b) also show a large reduction in the number of null hypotheses rejected in Karlan and List (2007) once multiple testing is taken into account. These examples serve to encourage researchers to identify and properly correct for all the different comparisons within a study to avoid the false positives that mechanically result from multiple testing.

Besides *ex post* corrections, researchers should pre-emptively take into account the problem of multiple hypothesis testing in the design phase. Intuitively, to control the false positive rate across all comparisons, stricter significance level requirements should be applied for each individual test *ex ante*. In practice, this means specifying lower levels of α in the power calculation for each comparison (see Section 6 for details on power calculations for single comparisons). Acknowledging this imperative already in the design phase reveals a "hidden cost" of adding another outcome, treatment arm, or subsample analysis to an experiment: every additional comparison the researcher plans to perform increases **in all existing comparisons** the number of participants (or the precision of measurement) that

[61] List et al. (2019b) made their MATLAB and Stata code available to other researchers for easy implementation of the procedure at `https://github.com/seidelj/mht`; see Seidel and Xu (2016) for documentation.

is required to maintain control over the study-level false positive rate. As a simple example, consider a researcher trying to determine the optimal sample size for an experiment that compares a treatment and a control group along two different outcomes. In order to ensure a study-level false positive rate of 5%, the researcher can use the above-described method by Bonferroni (1935) and set the significance level cut-off to $0.05/2 = 0.025$ for each individual comparison, and calculate the optimal sample sizes accordingly.[62]

Ensuring sufficient statistical power in a study while accounting for multiple comparisons may substantially increase the number of participants required and thus the cost of an experiment. As discussed in Section 6, appropriate design choices can be helpful in increasing statistical power without expanding the experimental budget. In the following two sections, we review in detail two such techniques that have the potential to reduce the variance of the estimated treatment effect: blocked randomization and within subject experimental designs.

8 Use blocked randomization to increase power and credibility

In Section 6, we have shown that the statistical power of a study is decreasing in the variance of the estimated treatment effect. We also highlighted approaches to reduce this variance by optimally choosing the *ratio* of subjects assigned to the treatment vs. the control group. This section considers more broadly the process of assigning treatment status to any given participant. In the following, we review the merits of blocked randomization compared to complete randomization in terms of statistical power. We then approach the topic from a different angle and show that blocking may reduce bias by serving as a

[62] Of course, this heuristic represents an overly conservative approach compared to our preferred MHT correction method byList et al. (2019b) that has more power than traditional approaches to correct for MHT.

commitment device against specification searching. Finally, we discuss the choice between **randomization** and **optimization**.

In the section on Preliminaries, we outlined the logic for randomly assigning subjects to treatments: randomization balances the treatment and control groups both in terms of observables and unobservables, allowing an unbiased identification of the treatment effect. However, in a completely randomized design, the variance of outcomes is potentially very large, and the sample sizes of treatment and control are randomly generated (List et al. 2011; Deaton and Cartwright 2018). As Duflo et al. (2007) point out, pure randomization only achieves balance *in expectation*: in practice, especially in the case of smaller samples, randomization may yield experimental groups that differ from each other along important observable dimensions. A popular way to address this issue is to include covariates in the estimation *ex post*. However, when data on subjects' relevant observable characteristics are available prior to conducting the experiment, it is preferable to use this information in the design phase and improve the overall precision of the study through **blocked randomization**.[63]

Blocking (also knows as stratification) refers to the practice of dividing experimental subjects into blocks (strata) by observable characteristics, such that randomization is performed within, but not between, these blocks (List et al. 2011). More formally, blocking involves partitioning the covariate space into a finite set and carrying out a completely randomized experiment within each of these subsets (Athey and Imbens 2017a). Using Neyman's repeated sampling approach, we can estimate the average treatment effect within

[63] An alternative approach to dealing with covariate imbalance is re-randomization (Morgan and Rubin 2012; Bruhn and McKenzie 2009; Banerjee et al. 2017b). Two commonly used forms of re-randomization are the "big stick" method that requires a new random draw if the imbalance between treatment and control groups in the resulting allocation exceeds a pre-specified threshold, and the "minimum maximum t-stat" method that suggests performing multiple (1,000 or 10,000) draws, checking for balance each time, then choosing the draw with the minimum maximum t-stat. Bruhn and McKenzie (2009) show that for very persistent outcome variables, and in smaller samples, blocked randomization performs better than re-randomization.

each block as the difference between the average outcomes for treated and control subjects, then estimate the overall average effect of the treatment by averaging the within-block estimates weighted by the share of subjects assigned to the block. In case the share of treated subjects is the same in each block, this simplifies to the difference in means between treated and control subjects – the same estimator we use for completely randomized designs (Athey and Imbens 2017a). Blocked randomization is beneficial because it increases precision: the estimated variance of the treatment effect is smaller once we take into account the gains from stratification. Compared to *ex post* regression adjustment, blocking in the design phase is preferred because it can ensure that the share of treated subjects is the same in each stratum, minimizing the variance of the estimate overall.[64]

Despite popular beliefs to the contrary, blocking does not lower precision *ex ante* even when the correlation between the outcome variable and the covariates on which we block is weak.[65] Note that the same is not true for *ex post* adjustments: adding covariates that do not explain the outcome variable in a regression increases standard errors by reducing the degrees of freedom (Duflo et al. 2007). Consequently, one should stratify on a rich set of covariates whenever possible, including continuous variables (Moore 2012).[66] The limit of stratified randomization is a paired design where each block contains only two observations: a treated and a control subject. While this approach has advantages in terms of precision, it complicates the subsequent estimation of variance (Athey and Imbens 2017a).

[64]Following Athey and Imbens (2017a), we obtain the estimated variance of the treatment effect with blocked randomization as follows (where g indexes blocks):

$$\hat{V}^{blocked} = \sum_{g=1}^{G} \hat{V}(\hat{\tau}_g) \left(\frac{N_g}{N}\right)^2 \quad \text{where} \quad \hat{\tau}_g = \bar{Y}_{T,g} - \bar{Y}_{C,g} \quad \text{and} \quad \hat{V}(\hat{\tau}_g) = \frac{\sigma^2_{C,g}}{n_{C,g}} + \frac{\sigma^2_{T,g}}{n_{T,g}}. \tag{7}$$

Comparing the expression in (7) with the variance estimate for completely randomized experiments presented in Section 6, $\sigma^2_T/n_T + \sigma^2_C/n_C$, in general we find that the latter is more conservative.

[65]See Athey and Imbens (2017a) for an explanation and two important qualifications to this result.

[66]A fascinating recent paper explores blocking on the predicted treatment effects and on subjects' willingness-to-pay for the treatments in order to design an "ethical experiment" that takes subjects' predicted welfare into consideration (Narita 2018).

Such a perfect matched pairs design also comes at high attrition costs for the matched units, and should thus be applied with caution.

Stratification is also desirable when the researcher expects heterogeneity in response to the treatment and wants to analyze subsamples separately. In this case, *ex ante* blocking maximizes power for estimating the treatment effect for each subsample. Equally importantly, stratifying on variables that the researcher *ex ante* deems as relevant increases the credibility of the study: it demonstrates to the reader that the subsample analysis presented in the research paper was actually planned in advance and is not merely the result of "data mining" or a "fishing expedition." In this sense, blocking on variables to be used in subsequent heterogeneity analysis helps address the problem of researcher bias discussed in Section 4, by limiting analytical flexibility (Munafò et al. 2017). If a researcher uses blocking primarily for credibility reasons rather than to increase precision, she should limit herself to blocking only on a few key variables. While hypothesis registries and pre-analysis plans (Christensen and Miguel 2018; Coffman and Niederle 2015) may provide stronger remedies against specification searching, blocking has the advantage of also increasing the power of the resulting subsample analyses. Note, however, that blocking alone does not address all the issues that arise from subgroup analysis: even if the experimenter "ties herself to the mast" by stratifying on the relevant dimensions, standard errors still need to be adjusted *ex post* for multiple hypothesis testing (see Section 7 for more details).

Utilizing baseline information to an even greater degree than in the stratification case, some researchers have recently suggested relying on **optimization** instead of randomization for assigning treatment status to subjects. Bertsimas et al. (2015) propose a method based on discrete linear optimization, such that assignment is chosen to minimize the discrepancy between treatment groups in terms of means and variances of covariates. Kasy (2016) considers the experiment as a statistical decision problem where the goal is to find the unique treatment assignment that minimizes a Bayesian or minimax risk function

(based on the mean squared error of a point estimator). While these approaches have the potential to increase power, gains in precision are substantial only when baseline variables strongly predict future outcomes (see Bruhn and McKenzie (2009) for an illustration of this point), and come at the cost of more complicated inference (a bootstrap method is required to obtain the p-values of the estimates).[67] Banerjee et al. (2017b) model the experimenter's problem in a Bayesian decision theoretical framework. They argue that for a given prior over treatment effects, there exists a deterministic treatment assignment that maximizes the experimenter's expected utility, leading to the proposition *"Bayesians do not Randomize."* Once multiple decision makers (or a single decision maker with different priors) are considered, however, they identify randomization as the only method to yield results whose interpretation cannot be challenged.

Overall, we are of the opinion that optimization on the basis of baseline covariates may be a useful method for assigning subjects to treatments in pilot studies. Sample sizes in pilots are typically small, so an increase in power is crucial. Furthermore, it is easily justifiable to design pilots so that they are most informative for a specific prior (that of the experimenter) rather than for a wider audience with arbitrary priors. For most experiments, randomization with "improvements" such as stratification or re-randomization remains more suitable.[68]

9 Use within subject designs when appropriate

In our overview so far, we have focused on experiments consisting of a single period where each subject is assigned to either the control or the treatment condition. These cases fall

[67] For more details, see David McKenzie's excellent blog post on the topic: `https://blogs.worldbank.org/impactevaluations/optimization-just-re-randomization-redux-thoughts-recent-don-t-randomize-optimize`

[68] See footnote 63 for details on re-randomization. As another possible design improvement, consider Wilhelm et al. (2017)'s procedure based on an orthogonal greedy algorithm that uses pre-experimental data to inform, rather than treatment assignment, the choice of both the sample size and the covariates to be collected.

into the category of **between subject (BS) designs**, because the estimated treatment effect is obtained through a comparison of means *between* the two groups. This represents the current state of art when economists generate data. Of course, researchers have the choice to collect data in multiple periods, allowing for the use of a **within subject (WS) design**, such that the same individual experiences different treatment conditions in subsequent periods. In the following, we discuss the benefits and threats associated with using within subject designs along the dimensions of statistical power, confoundedness, and heterogeneity in response to treatments.

Within subject designs have been advocated for their potential to yield **more powerful tests** for the same cost than between subject experiments (Charness et al. 2012). In particular, they allow for estimations controlling for individual-specific effects, reducing the variance of the treatment effect estimator to the extent that within-subject correlations explain the outcome (Frison and Pocock 1992; McKenzie 2012). Bellemare et al. (2016) suggest an approach based on Monte Carlo simulations to compare the power achieved in BS vs. WS designs. They provide a numeric example in the context of field experiments on gift exchange, and find that a BS design requires 4-8 times more subjects than a WS design to reach an acceptable level of statistical power (as discussed in Section 4, the conventionally required level of power is 80%). They also emphasize that adding more experimental periods can substantially increase the statistical power of a WS design, but has very little effect in the BS design.[69]

Note that in the above comparison we ignored cost considerations, assuming that collecting data from n subjects twice (WS design) is as costly as collecting data from $2n$ subjects (BS design). In practice, however, this is often not the case; in laboratory experi-

[69]While a within-subject design typically leads to lower variances for the estimated treatment effect, this is not necessarily the case. See Keren and Lewis (1993) for more details on precision in WS versus BS designs in the presence of treatment effects that are correlated with the individual-specific error term. For panel data with a non-i.i.d. error structures, we recommend Burlig et al. (2017)'s power calculation method that accounts for serial correlation in errors.

ments, adding additional periods to an experiment often comes at no additional monetary cost for the researchers (think of the typical practice of determining subjects' earnings based on their behavior in one period randomly selected at the end of the experiment). Field experiments, on the other hand, often have large per-period fixed costs (e.g. hiring and training surveyors) that make additional rounds of data collection more expensive on the margin.

Despite the fact that within subject designs have the potential to achieve better precision, in practice researchers do not seem to take the choice of design into account when setting the number of experimental subjects: surveying two recent volumes of the journal *Experimental Economics*, Bellemare et al. (2014) find that the median number of participants per treatment was relatively similar (43.5 and 50, respectively) for studies using BS and WS designs. They also find that the majority of studies in their survey (41 out of 58) are based on BS experimental designs. The relative unpopularity of within subject designs may be due to the strong assumption they require for inference. When the same subject is exposed to different treatment conditions, within-subject comparisons only provide a causal estimate if there is **independence of these multiple exposures** (Charness et al. 2012). There are different reasons why this assumption may not hold, such as learning, history and demand effects, and sensitization to perceived dependencies between treatments (List et al. 2011; Keren and Lewis 1993).

As a result, findings from WS designs may be confounded and hard to interpret. Crossover designs (where subjects are exposed to treatments in random order) may ameliorate, but not eliminate, these fears. The extent to which confoundedness is a threat depends very much on the particular research question. In some cases, one design is clearly better suited to test a particular theory (consider, for instance, predictions about individual preference reversals). When treatments are suspected of having persistent psychological effects or when experimenter demand effect is a concern, researchers should be cautious when us-

ing WS designs.[70] On the other hand, skill-based experiments, such as the study of Smith et al. (1989) on eyewitness accuracy and confidence, are less likely to yield a biased result under a WS design (Charness et al. 2012). WS designs may also work better in cases when the first treatment corresponds to the "status quo" (i.e. it conforms to the set of beliefs or expectations participants have already held coming in to the experiment) such that it does not induce learning nor cause sensitization to some aspect of the treatment.[71]

This trade-off between power and bias has been a central theme of the within versus between subject debate. We would like to conclude this section by emphasizing another aspect of the design choice that often receives less attention: between and within subject designs differ in the extent to which they are informative of individual differences in response to treatment. As we mentioned in the Preliminaries, a between subject comparison with random assignment to treatment only produces an unbiased estimate of the *average* treatment effect, but does not identify other moments of the distribution. While the average treatment effect conveys important information regarding the existence of an association or effect, it may mask crucial differences between participants in their reaction to treatment. For instance, assessing the share of the population that was helped or harmed by the treatment requires knowledge of the distribution of the difference between the outcomes of each individual in the presence and absence of the program.

In experiments using between-subject treatment assignment, each participant is only observed in one of the states, thus welfare analysis requires additional, often strong, assumptions.[72] Within person designs facilitate welfare calculations by measuring baseline

[70] See Bohnet et al. (2016) for a discussion on preference reversals between separate and joint evaluations and Hsee (1996) for an overview of the literature on evaluability.

[71] An example where within and between subject designs yielded very similar conclusions comes from two laboratory experiments studying the impact of affirmative action policies on participants' willingness to compete. In a between subject design, Balafoutas and Sutter (2012) find the same result as obtained by Niederle et al. (2013) in a within subject design that gender quotas induce high-performing women to enter tournaments without discouraging men from competing.

[72] Quantile regressions, for example, are only informative of the distribution of individual changes in outcomes if a rank invariance condition is satisfied. Bedoya Arguelles et al. (2018) provide an excellent

outcomes together with changes in response to the treatments for the different baseline values, providing the entire joint distribution rather than marginals. Allcott and Taubinsky (2015) use a within-subject design in their information nudge experiment to estimate the average change in valuation induced by their nudge *for each level of initial valuation*. This strategy allows them to calculate the market demand curve and the average marginal bias, statistics they show are sufficient for computing the welfare effects of a policy.

Further, WS designs can help distinguish between behavioral theories by showing heterogeneity in preferences *within* individuals. Gibson et al. (2013), for instance, refute the type-based model of lying aversion by documenting differences within individuals (across situations) in the estimated cost of lying.

WS designs can also help researchers uncover heterogeneous treatment effects. Data from WS experiments can be used to plot a histogram of the realized "individual linear differences," calculated for each subject as the difference between their outcome in the treated vs. the control state. Such histograms may hint at important dimensions of heterogeneity, and could suggest subgroups that benefit most/least from the treatment.[73] As an example, consider the study of Hoel (2015) on the role of asymmetric information in dictator games. While she finds that subjects on average give more in games when the choice is public than when it is secret, a within subject comparison reveals that almost half of the participants give the same amount in both conditions. Her design allows her to classify participants into types based on their individual response to the treatment, such that types identified in the laboratory also behave differently in the field.

An alternative approach to studying heterogeneous treatment effects that does not require multiple observations per participants makes use of machine learning (ML) tech-

summary of different methods aimed at identifying the distribution of individual specific treatment effects in RCTs.

[73]It is important to emphasize that WS designs on their own still do not allow *identification* of the distribution of treatment effects. Readers interested in identifying the probability distribution of individual effects in panel data should consult e.g. Arellano and Bonhomme (2012).

niques such as *causal trees* (Athey and Imbens 2016) and *causal forests* (Wager and Athey 2018). As discussed in Section 7, this approach discovers heterogeneity by seeking to partition the data into subgroups with different treatment effects (Athey and Imbens 2017b). While applications of ML techniques have considerable promise for predicting treatment response differences based on observable covariates, there are two issues we need to keep in mind. First, ensuring the consistency of estimates requires an "honest" approach to estimation, whereby the sample is split to ensure that the data used for partitioning the covariate space are different from data used for estimation of the treatment effects. Testing the out-of-sample accuracy of predictions also requires a hold-out sample that is not used at all for training or estimation. As such, these methods tend to work best with larger data sets (Davis and Heller 2017). Second, in case some of the covariates are correlated with each other, consistency in model selection is not ensured: the importance of specific covariates for prediction may vary across sample partitions (Mullainathan and Spiess 2017). As such, these techniques do not allow the researcher to conclude that certain observables are not associated with treatment effect heterogeneity just because they were not used by the algorithm to create the prediction.

In sum, we encourage researchers to carefully weigh the pros and cons of between and within subject designs along the dimensions discussed above (power, cost, bias, learning about heterogeneity) and pick the one better suited to answering their particular research question. Just as in the case of the choice between lab or field experiments discussed in Section 3, there is no universally preferred method: the choice to vary treatment conditions within or between subjects should depend on the characteristics of the experiment and the nature of what information is sought and the trade-offs the researcher is willing to make. In a nutshell, it depends on the theory to be tested – a topic we discuss in the next section.

10 Go beyond A/B testing by using theoretically-guided designs

In Section 6 and the discussion of optimal experimental design that followed, we mainly focused on experiments whose main goal was to measure whether one treatment condition yields a different mean outcome than another treatment and/or the control condition. A simple example for this approach is "A/B testing", a method common both in research and in business, that entails showing subjects one of two versions of the same product or service at random, and comparing responses. Yet, the experimental method will never reach its true potential unless we go beyond simply documenting the existence/size of an effect. Rather, we should exploits experiments' ability to generate data that allows us to explore the underlying mechanisms at work, to understand the **whys** behind the data patterns observed. We therefore need to design experiments tightly linked to economic theory in order to reap the true benefits of the experimental method.

Despite its advantages, using theory is still not a commonplace occurrence: out of all the experiments published in the top 5 journals in Economics (*American Economic Review, Econometrica, Journal of Political Economy, the Quarterly Journal of Economics, Review of Economic Studies*) between 1975 and 2010, 68% were "descriptive", meaning that they lacked an economic model; of the remaining articles, only 14% allowed for more than a single model, either by directly comparing competing models or by estimating one or more structural parameters (Card et al. 2011). The theory-free nature of RCTs is a serious disadvantage in attempting to generalize (Deaton and Cartwright 2018). By combining experiments with theory we can reap the best of both worlds: preserving causal inference (due to the exogenous variation created by the experiment) and improving the generalizability of predictions through theory and structural estimation (for a recent example see DellaVigna et al. 2012).

To address this gap, we advocate for experimental economists to use economic theory to *inform their experimental design* whenever possible (Banerjee 2005; Heckman 2010; List 2011), and to incorporate results from experiments into existing economic theory (Deaton and Cartwright 2018), thereby creating a feedback process that guides the development of theory and the design of future experiments (Duflo et al. 2007). Combining economic theory with experiments allows researchers to estimate a wider array of parameters (Attanasio and Meghir 2012), to perform counterfactual analysis, to test theories directly (Browning and Chiappori 1998), and to account for general equilibrium and welfare effects. In the following, we review in detail the benefits from designing experiments in connection with economic theory.

First, we can use theory to explicitly model **selection** into the experiment. When running lab experiments, AFE, and FFE, researchers should make explicit their assumptions about selection by using theory and, in doing so, address some of the concerns about generalizability discussed in Section 1. For instance, researchers could specify a variant of the Roy model to describe selection into the experiment (Heckman 2010).

Using economic theory jointly with an experiment allows the researcher to perform **structural estimation** to estimate the parameters of a theoretical model from data collected in an experiment. This enables researchers to perform **ex-ante counterfactual policy analysis**: to predict the effects of policies or programs that have not yet been implemented and over which no data are available (Heckman 2010).[74] Therefore, economic theory allows researchers to extrapolate the results of existing experiments to other populations and settings (Banerjee 2005; Falk and Heckman 2009, see Section 12 for a discussion on scalability).

Coupling experiments with structural estimation allows researchers to understand the **mechanisms** underlying the observed behavior in the experiment. As an example, con-

[74]For a discussion of structural estimation and reduced-form estimation, see Nevo and Whinston (2010).

sider the RCT by Dupas (2014) that employed a two-stage randomization of the pricing of a novel product to differentiate between two possible mechanisms of policy adoption; or the field experiment by Chandrasekhar et al. (2018) designed to separate two mechanisms (reputation vs. shame) in a model of stigma. Moreover, researchers can design their experiment with the structural model in mind, in such a way that makes the **identification of the relevant parameters** possible. For example, as part of their experiment, Hedblom et al. (2016) set up a "firm" to hire workers, and by varying the wages they paid as well as the level of corporate social responsibility perceived by their workers, were able to identify and estimate a structural model of unobserved worker heterogeneity. The identification of the relevant parameters also allows for measuring **welfare effects**. In a door-to-door donation experiment by DellaVigna et al. (2012), some households were warned of the solicitors' upcoming visit such that they could sort in or out of the intervention, and this sorting behavior (together with the donation choices of those who opened the door) allowed the identification of altruism and social pressure parameters, and of welfare estimates.

When researchers want to estimate a structural model, they can think about the "ideal data" for the identification of those parameters, and then design an experiment that generates exactly that type of data. But even when the experiment is already underway, researchers can still identify a relevant structural model using precisely the unique features of the experimental data. Low and Meghir (2017) discuss an interesting comparison between different approaches for using structural models to exploit experimental data in the context of PROGRESA, a conditional cash transfer program in Mexico aimed at increasing school participation in poor rural areas. Todd and Wolpin (2006) estimate a structural model of school participation, taking into account the opportunity cost for children; their focus is on the **validation** of their model (identified from control data only), that can then be used to make predictions. In contrast, Attanasio and Meghir (2012) estimate a model in which the grant can have a different marginal utility than other sources of in-

come (such as child wages), what allows for identification of the effect of the grant even in the presence of general equilibrium effects. As these examples demonstrate, the different advantages of combining structural estimation with experiments we have surveyed are not mutually exclusive, and researchers can find creative ways to exploit these synergies.

Economic theory is also useful when considering **spillover effects**, whereby subjects impose externalities on others (for example, those in treatment can inform their friends in the control), and **general equilibrium (GE) effects**, that occur when agents react to the intervention in a way that changes the environment itself (Duflo et al. 2007; Maniadis et al. 2015).[75] GE effects are of great importance in diverse contexts ranging from economic development (Banerjee 2005; Acemoglu 2010), to health care utilization (Finkelstein 2007) and the microcredit literature (Burke et al. 2014). However, most experiments are conducted assuming *partial equilibrium*, i.e. assuming away spillover and GE effects. Disregarding the GE effects of experiments might be justified for small interventions that only affect a few participants or that have a small impact, such as in lab experiments and artefactual field experiments (AFE). However, field experiments (FFE and NFE) can induce changes in the local economy that could translate into general equilibrium effects. Neglecting these GE effects biases the results and leads to misleading conclusions about the true effect of the intervention. This is especially true for large-scale interventions claiming to have a large impact. Such experiments should attempt to include measuring the general equilibrium effects as part of their experimental design whenever possible.

The measurement of GE effects typically requires an experimental design that explicitly accounts for them. For example, Crépon et al. (2013) included two levels of randomization in their experiment, one of them being the proportion of individuals assigned to treatment, allowing them to capture the GE effects of their intervention. Another exam-

[75]Note that spillover effects violate the "Stable Unit Treatment Value Assumption" (Angrist et al. 1996; Duflo et al. 2007) discussed in the Preliminaries. For the rest of the section, we will use the term "general equilibrium effects" to also include spillovers.

ple is the study by Cunha et al. (2017), who estimate the GE effects of cash versus in-kind transfers in Mexican villages: both types of transfers had similar value (allowing a partial equilibrium comparison), but the in-kind transfers also generated more supply of certain goods in the market, affecting the calculation of the intervention's general equilibrium effect.

The importance of measuring a program's **welfare effects** has been widely acknowledged (Heckman 2010). Recently, welfare analysis has made its way into behavioral economics, yielding a fruitful collaboration between economic theory and experiments: in addition to measuring the traditional outcomes of the experiment, researchers can use theory to infer the change in subjects' well-being as a consequence of treatment. DellaVigna et al. (2012) provide an early example of measuring welfare effects in a natural field experiment via structural estimation: they find that a door-to-door campaign of charity donation decreased the welfare of the average household due to the social pressure associated with not donating.[76]

However, welfare calculations may be sensitive to certain details. First, the particular theory researchers base their calculations on has important consequences for the conclusions on welfare (Jimenez-Gomez 2018). This aspect is even more salient in areas where individuals are known to be subject to behavioral biases.[77] Second, GE effects should be included in welfare calculations. As an illustration, consider Handel (2013)'s study of health insurance markets where substantial inertia had been observed. He structurally es-

[76]Related studies test specific aspects of behavioral theory and their welfare consequences through field experiments: (Zhe Jin et al. 2010; Bernheim et al. 2011; Allcott and Taubinsky 2015; Allcott and Kessler 2019; DellaVigna et al. 2017, 2016), and develop theories that explicitly consider how behavioral economics affects welfare measurement: (Spiegler 2014; Gabaix and Farhi 2017; Jimenez-Gomez 2017). Finkelstein and Notowidigdo (2018) use a randomized natural field experiment to test two competing explanations – based on neoclassical and behavioral theory, respectively – for the low take-up of SNAP benefits, and estimate the welfare impact of different interventions aimed at increasing take-up.

[77]When individuals face behavioral biases, welfare calculations using only the demand curve (and therefore ignoring those biases) can be potentially mistaken by orders of magnitude, and even have the wrong sign (Baicker et al. 2015).

timates a choice model in order to compute the welfare effects of a counterfactual "nudge reminder", and concludes that although the nudge would increase the rate at which people select into plans that better match their needs (reducing inertia), the GE effects would exacerbate adverse selection, leading to a price increase and a loss in average welfare. Finally, researchers should pay attention to potential heterogeneity in treatment response (see Sections 1 and 12). Heterogeneity is crucial when computing welfare effects, because different individuals may benefit (or suffer) to varying degrees from a given program, and therefore the distribution of welfare can be very different for some subpopulations (Jimenez-Gomez 2017).

Using economic theory is not without caveats. It is always possible that the particular economic theory we consider is wrong, and this concern has been exacerbated with the rise in importance of behavioral economics (Banerjee 2005). Moreover, structural estimates are sensitive to assumptions about functional forms and distribution of unobservables (Heckman 2010). The correct design of the experiment can never be undermined due to confidence in the theory.[78] To conclude, we would like to emphasize what we are **not** advocating. We do not call for every experiment to be derived from economic theory or to be structurally estimated. There are occasions when a descriptive study is perfectly appropriate, for example when attempting to differentiate between several competing theories whose predictions go in opposite directions (Acemoglu 2010). We also do not advocate for journals to demand that authors include ad-hoc economic models after the experiment has been conducted and the data analyzed. Such models add little value in our opinion and can confuse readers as to the true intent and nature of the studies. We **do** believe that there is value in descriptive experiments, but the limitations of these types of studies should be explicitly acknowledged. We also believe that in order to make generalizable

[78] Card et al. (2011) claim this was the case in the negative income tax experiments conducted in the late 1960s and early 1970s.

predictions, using economic theory to design experiments and to guide the analysis is often the correct choice (Heckman 2010; Acemoglu 2010).

11 Focus on the long run, not just on the short run

Economic experiments often tend to focus on estimating short-term substitution effects. This is probably due to the fact that conducting an experiment that follows up subjects for several months or years is substantially more costly: the logistics required become complex, and there is a need to incentivize subjects to come back for follow-ups to avoid attrition.[79] In addition, there is always an implicit opportunity cost associated with longer experiments, because their longer time to completion delays publication compared to similar trials that focus on short-term effects.

However, understanding the long-term effects of interventions is critical. For example, demand elasticities can be very different in the short-run vs. the long-run (Huang et al. 2017). Long-term effects are especially relevant when the interventions are programs that governments or other organizations intend to roll out in large scale (we discuss scalability in Section 12). As we emphasized in Section 10, it is fundamental to take into account the general equilibrium (GE) effects of the implemented policies and programs. However, those GE effects often need time to manifest, so measuring the long-term effect of interventions is even more important. Moreover, Hawthorne, John Henry, and experimenter demand effects can in principle be identified by collecting long-run data (Duflo et al. 2007). In addition, the return on investment (ROI) per dollar will be much larger if the effects persist in the long-term. For example, Levitt et al. (2016) provided financial incentives to high school freshmen for 8 months conditional on meeting an achievement standard, and followed the participants for five years. They found a significant positive impact of the in-

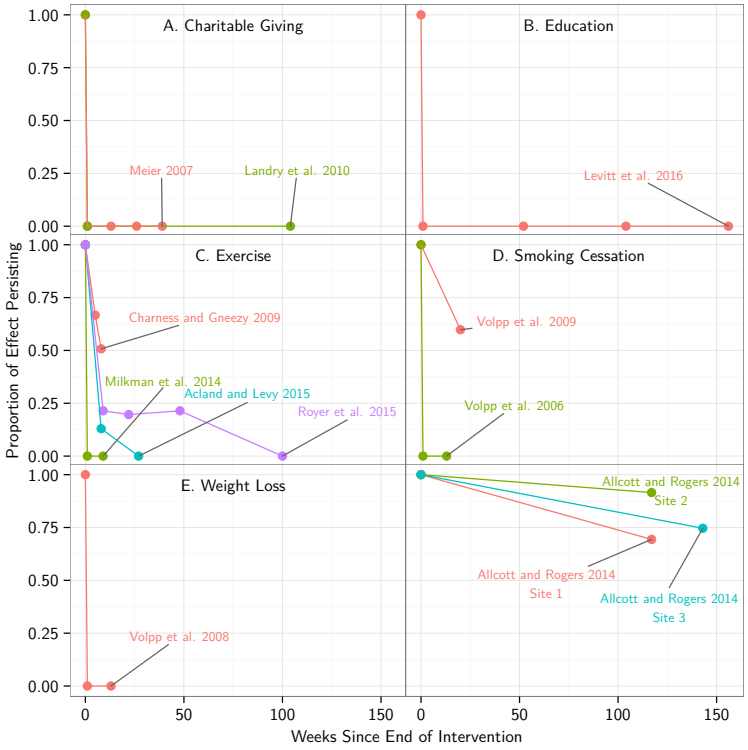
[79] For example, Charness and Gneezy (2009) paid 50 US dollars to subjects for each of two follow-ups.

centives on academic achievement in the short run, but the long-term follow-up revealed that the gains did not persist beyond the first year.

Despite the widespread focus on the short term in experimental economics, there are several notable papers that analyze the medium- and long-run effects of experimental interventions.[80] As Brandon et al. (2017) demonstrate, the evidence on whether interventions work in the long run is mixed even when restricted to a single subfield. Out of ten studies included in their review (covering the fields of charitable giving, education, exercise, smoking cessation and weight loss), four are consistent with habit formation, whereas the other six are not (see Figure 2, plotting the proportion of the intervention's effect that persists after incentives are removed). Moreover, even when the effects persist, they decay rapidly: only two of the aforementioned studies found estimated effect sizes larger than 25 percent of the initial effect after just a month, and only one found any persistence after six months.

There are two potential and non-excluding reasons why researchers may find no or small effects of interventions in the long run, with very different implications for the actual existence of a treatment effect. The first and most obvious explanation is that the effect of the interventions is truly zero in the long run. This could happen if the intervention is not successful in changing behavior in the first place, either because subjects do not enroll in the relevant treatment, because the incentives or nudges are "too weak," or because the treatment changes behavior in a way different than expected. Even if the treatment is suc-

[80] We found relevant papers in the following areas: exercise (Charness and Gneezy 2009; Milkman et al. 2013; Royer et al. 2015; Acland and Levy 2015), smoking cessation (Volpp et al. 2006, 2009; Giné et al. 2010), weight loss (Volpp et al. 2008; Burke et al. 2012; John et al. 2011; Anderson et al. 2009), charitable giving (Meier 2007; Shang and Croson 2009; Landry et al. 2010), water conservation (Ferraro et al. 2011; Ferraro and Price 2013), energy conservation (Allcott and Rogers 2014; Brandon et al. 2017), voting (Gerber et al. 2003), labor effort (Gneezy and List 2006), exposure to better neighborhoods (Chetty et al. 2016) and education (Jackson 2010; Jensen 2010; Walton and Cohen 2011; Rodriguez-Planas 2012; Levitt et al. 2016). Note that some of these papers would be better classified as belonging to the literatures on psychology, medicine and behavior change, but are included here for completeness. In their review, Rogers and Frey (2016) analyze how field interventions to improve societal well[U+2010]being work over time.



Reproduced with permission from Brandon et al. (2017): "each point represents the proportion of the initial treatment that persists for a given amount of time since the end of a given intervention [...] insignificance at the five percent level constituting persistence of zero."

Figure 2: Persistence of effects across domains

cessful in changing behavior in the short-term, subjects could revert to baseline behavior over time even in the presence of the intervention: as the novelty of the incentives tapers off, or if the intervention crowds out intrinsic motivation, subjects revert to old habits.

Yet, there is a second reason why long-run estimates of treatment effects are so often (close to) zero: the **"attenuation bias over time."** Consider a field experiment where subjects are assigned to a treatment $z_i \in Z$. The researchers would like to measure the Average Treatment Effect at time t, $\tau_t^* = \mathbb{E}[(y_{i1t} - y_{i1t_0}) - (y_{i0t} - y_{i0t_0})]$.[81] Note, however, that they can only measure $\tau = \mathbb{E}[y_{it} - y_{it_0}|z_{i0} = 1] - \mathbb{E}[y_{it} - y_{it_0}|z_{i0} = 0]$.[82] Importantly, z_{i0}

[81] Where y_{idt} is the outcome of individual i in treatment d at time t, and t_0 is the time at which the intervention starts.

[82] Where $y_{it} = y_{id_{it}t}$ is the outcome of individual i at time t, given the fact that individual i is in treat-

refers to the original assignment at time t_0, but subjects may change their behavior over time, effectively changing their treatment (we discuss this form of non-compliance in Section 1). Subjects may self-select into different treatments over time: we can think of this change as probabilistic, happening with a higher probability when the gains from changing treatment status are high and the cost of changing is low. As time goes by, attenuation bias increases and as time goes to infinity, the estimated average treatment effect approaches zero.[83]

If selection into the experiment ($p_i = 1$) is positively correlated with the utility from, or negatively correlated with the cost of changing treatments, then the attenuation bias will be exacerbated in overt experiments compared to natural field experiments. When researchers find an ATE close to zero due to attenuation bias over time in a an overt experiment, it does not mean that the treatment did not work in the first place, but that subjects who selected into the experiment ($p_i = 1$) found a way to improve their outcomes over time, even when they were initially assigned to control.[84] A solution to attenuation bias over time is to measure the effect of the intervention with respect to a control group that did not self-select into the experiment ($p = 0$): $\tau' = \mathbb{E}[y_{it} - y_{it_0}|p = 1, z_0 = 1] - \mathbb{E}[y_{it} - y_{t_0}|p = 0]$, where $y_{it} = y_{id_{it}t}$ is the outcome for individual i at time t. Note that τ' controls for changes that happen in the treated group, with respect to a group that is not part of the experiment.[85] This also allows researchers to identify cases when the treat-

ment d_{it}.

[83] In the Appendix, we assume that the occurrence of opportunities to change treatment follows a Poisson distribution, and that the actual change of treatment follows a Markov process, and formally show that in the limit, $\tau_t \to 0$.

[84] This could happen if those who selected into the experiment were more motivated, and hence more likely to find ways to enroll into alternative programs outside of the experiment that would improve their outcomes, or if there were *spillover effects* in the experiment (see Section 10) and those initially assigned to the control group were more likely to be affected by the treatment because of the fact that the experiment was being run (for example, their friends in the treatment group informed them about the ways to get enrolled in similar programs, etc).

[85] This can be seen because τ' gives us the correct measurement in the extreme cases when either the treatment has no effect but selection into the experiment has an effect (in which case $\tau' = 0$), or when selection into the experiment has no effect but the treatment has an effect (in which case $\tau' = \tau^*$).

ment works and has spillover effects, which could otherwise be mistaken for a lack of long-run effect.

For all the reasons mentioned above, studying long-run effects should be a routine practice of experimental economics whenever doing so enhances our scientific understanding of the theory, mechanisms, or facts around the question. To reduce the delay in publishing that long-term studies require, researchers could in principle continue to collect data after their first, short-term results have been published (Banerjee 2005). Moreover, it is often possible to track subjects over time without incurring additional costs, for example when data is already being collected for administrative purposes, or when researchers establish long-term collaborations with firms who continue to share their data beyond the initial experiment.

12 Understand the science of scaling ex ante and ex post

Throughout this paper, we have discussed problems related to statistical inference, generalization and reproducibility. These issues become especially salient when researchers attempt to **scale up** their interventions, i.e. to extend them to a population which is larger (and usually more diverse) than the original one. Unfortunately, out of the large number of program evaluations performed today, few programs are ever scaled, and when they are, the effect sizes often diminish substantially: a phenomenon known as "voltage drop" (Al-Ubaydli et al. 2017a). There are many possible reasons why such a voltage drop might occur, and it is crucial to understand why scaled-up programs often do not work as intended. Yet we believe that the problem is currently too narrowly defined along two important dimensions (Al-Ubaydli et al. 2017c). First, the discussion around voltage effects in the implementation science literature tends to focus mostly on the scaled-up program's benefits.

However, understanding the relative benefits and costs are both invaluable to the scalability discussion. Second, whereas that literature tends to focus on program fidelity as a major reason for the lack of proper scaling, we see three main areas where challenges to scalability arise: statistical inference, representativeness of the population and representativeness of the situation. Al-Ubaydli et al. (2017c) provide a formal model of the way these three factors manifest in the market for scientific knowledge; we simply sketch them below to highlight issues experimenters should consider in their design, analysis, and interpretation that could affect the scalability of their results.

Statistical inference. In Section 4, we introduced the concept of the post-study probability (PSP, Maniadis et al. 2014): the probability that a declaration of a research finding made upon statistical significance would actually be true. We discussed how insufficient statistical power combined with bias resulting from specification searching may lead to low post-study probabilities. We further explained that as more researchers investigate the same relationship independently of each other, the probability that a statistically significant effect or association truly exists becomes smaller. Moreover, studies with "surprising" results (i.e. low prior probabilities of being true) tend to be published in journals more often, in turn resulting in low PSP. For all these reasons, a program that is selected on the basis of just one successful initial trial may fail to produce effects when scaled. Furthermore, the phenomenon of effect inflation, i.e. obtaining an exaggerated estimate of the magnitude of a true effect (discussed in Section 4), implies that a program that is scaled after a single study is likely to yield a much smaller effect than the original study.

To curb these problems, Al-Ubaydli et al. (2017a) advocate for only advancing results to the policymaking stage once their PSP passes 95%. Crucially, the PSP increases substantially if the initial positive finding is followed by at least two successful replications (Section 5). Moreover, successful replications in different contexts are valuable for ensuring generalizability (Duflo 2004; Muralidharan and Niehaus 2017, see Section 1). Replica-

tion may also be used to measure average within- and across-study variation in outcomes: when these are close, the concern about across-context generalizability is reduced (Vivalt 2017).

Representativeness of the population. Heterogeneity in populations may present problems for scalability, as discussed in the context of generalizability in Section 1. During scaling, the population that selects into the program is often different than the original experimental sample, raising the concern that the estimated treatment effect will be different (usually smaller) in the new population. Such "scaling bias" may result from *adverse heterogeneity* (Al-Ubaydli et al. 2017c), describing a situation when the original experimental participants' attributes are correlated with higher expected outcomes. This may occur as a result of participation bias (participants self-select into the experiment on the basis of their expected gains from participation Al-Ubaydli and List 2013) or publication bias (researchers have incentives to find participants who yield large treatment effects Al-Ubaydli et al. 2017c). Another concern for scalability related to the population is attrition, an issue we discussed in detail in Section 1.

Representativeness of the situation. The first and most obvious change when scaling up a smaller program concerns the infrastructure: scaled-up programs often need a larger and more complex infrastructure to support them, leading to a potential increase in their cost. Moreover, program evaluations are often run by particularly high-quality officials or NGOs in a way that is hardly possible to scale up: as the program is expanded, its quality may deteriorate (Duflo et al. 2007). When scaling, researchers need to be aware that more workers must be hired, and they might be of lower quality or have less interest in the program, simply due to the diseconomies of scale associated with requiring more labor in a labor market with inelastically-supplied human capital (Al-Ubaydli et al. 2017c; Davis et al. 2017; Banerjee et al. 2017a). This may reduce the benefits of the program (through lower-quality program workers) and/or increase its costs (when trying to keep

the quality of workers high). In addition, the cost of recruiting more subjects may be lower or higher than in the initial experiment, depending on the particular implementation of the scaled-up program (for example, a program that automatically enrolls citizens can have lower marginal costs, once the setup for recruitment is ready).[86]

In addition to the three areas listed above, **general equilibrium (GE) effects** are also crucial when considering scaling (Banerjee et al. 2017b; Al-Ubaydli et al. 2017b, also Section 10 for the importance of GE effects). GE effects can cause researchers to underestimate the effect of their interventions in at least two ways: if there are spillovers from the treated group to the control group, in such a way that the ATE was biased downwards in the original intervention because the control group was benefiting from treatment; and if there are complementarities between those treated, for example in an educational intervention in which students benefit not only from their own treatment, but also from having treated peers.[87]

On the other hand, researchers may overestimate the effect of their intervention if they do not take into account its crowding out effect (Crépon et al. 2013).[88] Moreover, the interventions that work when scaling up might be more complex than those employed while piloting the program.For example, in the case of long-term, chronic medical conditions, the

[86]Interestingly, many of these issues could be exacerbated "when rolling out revolutionary ideas, as these often challenge the power and established practices of incumbent organizations" (Al-Ubaydli et al. 2017c). Therefore, programs with greater community coalition functions, communication to stakeholders, and sustainability are more likely to still be in place over two or more years beyond their original funding (Cooper et al. 2015).

[87]Dramatic evidence of such spillovers comes from List et al. (2019a), who examine a randomized field experiment among 3-5 year olds in Chicago described in Fryer et al. (2015, 2017). They find that each additional treated child residing within a three kilometer radius of a control child's home increases that child's cognitive score by 0.0033 to 0.0042 standard deviations. Given that an average child in their sample has 178 treated neighbors residing within a three-kilometer radius of her home, on average, a child gains between 0.6 to 0.7 in cognitive test scores and about 1.2 in non-cognitive test scores in spillover effects from her treated neighbors. These are large spillovers, which serve to highlight that the program at scale would have much larger effects than the Fryer et al. (2015, 2017) summaries of the research program predicted, ceteris paribus.

[88]A related issue is that of *construal*, i.e. the subjects' subjective understanding of the intervention. Paluck and Shafir (2017) argue that scaling up mandatory arrest of abusive domestic partners (as a result of an experiment) backfired due to the construal of (what it meant) calling the police in that situation.

most effective interventions are usually complex combinations of basic interventions such as educational sessions, counseling, and a selection of reminder methods (Al-Ubaydli et al. 2017a). Therefore, the benefit of a scaled-up program can be higher or lower than that of the original intervention, depending on the direction of the GE effects.

In light of the scaling problem, our main recommendation is that researchers "**backward induct**", having the issue of scaling already in mind when designing their experiments and programs. First, both clinical researchers and economists can greatly benefit from following experimental best practices, such as emphasizing sound inference (Sections 4 and 7), ensuring appropriate sample sizes and sufficient power (Section 6) and conducting replications (Section 5). Moreover, a unified framework for addressing scalability should consider the underlying mechanisms of the program, as well as the relevant population, time-span, implementation partners needed, etc. In short, programs are more likely to scale up when they are driven by an understanding of the underlying mechanisms, the randomization (of subjects and of workers) happens in a large enough population, and the time scale is long enough to accurately measure the main effects as well as spillovers and GE effects.

First, in terms of mechanisms, researchers should go beyond A/B testing to the whys of the phenomena they want to study, using existing evidence to build theories that can help explain the experimental results, and providing a theoretical basis for fidelity in the Al-Ubaydli et al. (2017c) model. Economists could structurally estimate behavioral models as they seek to scale results, as we argued in Section 10.

Researchers should also consider whether results from their program are likely to generalize, being especially sensitive to heterogeneity across populations and contexts (Section 1), and choose the optimal experiment type in light of their scaling goals (an issue we discussed in Section 3). In terms of the population, an approach that is likely to result in better generalization is to consider the (large) population of interest first, take a representative (smaller) sample of observations this population, and then randomize those to treat-

ment or control (Muralidharan and Niehaus 2017).[89] Compliance to the program must also be taken into account. Lessons derived from earlier interventions can help in this endeavor: Al-Ubaydli et al. (2017a) argue that there is much to be learned from medical researchers who have been rigorously studying a similar problem for years.[90]

The issue of diseconomies of scaling in hiring workers to scale up an intervention can be tackled by randomizing the hiring process itself as part of the initial experiment, in a way that allows researchers to estimate the loss in worker quality as they move along the supply curve (Davis et al. 2017). Researchers can then use these insights in their initial cost effectiveness calculations before rolling out the experiment on a larger scale. In addition, GE effects can be estimated by using large units of randomization whenever possible (for example, randomizing at the school level instead of at the student level, Muralidharan and Niehaus 2017).[91]

Finally, it is important to evaluate the program's implementation, and document all steps of the process (Duflo et al. 2007; Banerjee et al. 2017a). This includes having a pre-analysis plan, and creating an initial program that is modular, in the sense that its implementation can described by a simple protocol (Banerjee 2005).

It is worth adding that machine learning can offer new opportunities for improving the scaling of interventions. Machine learning offers several potential advantages that economists can profit from (Mullainathan and Spiess 2017; Athey 2018), such as providing a benchmark against which to test economic theory, predicting who will benefit from a certain policy (for example, Björkegren and Grissen 2018, use machine learning to predict loan

[89]For example, Muralidharan and Sundararaman (2015) first sample a "representative universe" of villages with a private school, and then randomly assign each of them to treatment or control in a school choice experiment.

[90]In particular, non-adherence to medication can lead to financial and personal costs – incentives that nonetheless seem too weak to motivate individuals.

[91]A prominent example is that of Miguel and Kremer (2004) who realized the importance of spillover effects from deworming programs by randomizing the programs at the school (rather than the student) level. However, Banerjee et al. (2017a) warn that sometimes it is difficult to know ex ante what randomization unit will be large enough to capture all GE effects.

repayment using cellphone data),[92] and estimating heterogeneous treatment effects. While machine learning and artificial intelligence have only started making their way into economic research, they have great potential and, when combined with experiments, offer the promise of improving the scaling of interventions.

Conclusion

When new areas of inquiry arise in the sciences, they are oftentimes greeted with much skepticism, yet if they prove fruitful, they grow quickly but many times non-optimally. Such an oft-observed pattern effectively results in a missed moment to advance knowledge, and backtracking on ill-advised research journeys often is difficult. We now find ourselves in a phase of rapid growth of the experimental method in economics, especially as applied in the field. To help ensure that we seize this opportunity to maximize the scientific knowledge created using experiments, we take this occasion to step back and attempt to set up some guard rails by crafting a 12 item wish list that we hope scholars can do more of in their own research. By creating such a list we are not implying that these 12 items are entirely ignored in the extant literature; indeed, throughout our paper we highlight examples of research that already engages in these best practices.

While picking a dozen items for such an exercise is akin to picking one's favorite research project or one's favorite child, we nevertheless attempt in this tome to do just that. Rather than regurgitate how our twelve items span three bins that are usefully summarized by three questions, we wish to close with a few items that we find important but just missed our wish list.

Our first addition is a call to define the research questions and analysis plan before

[92] Other examples of machine learning for predicting behavior are Glaeser et al. (2016), who crowdsourced an algorithm for predicting health code violations in restaurants; and Goel et al. (2016), who predict the likelihood that a target of stop-and-frisk policies actually has a weapon.

observing the outcomes of an experiment – a practice known as preregistration (Nosek et al. 2018). This approach has been adopted as a remedy against specification searching in a plethora of other fields, most notably medical trials, and we fully anticipate considerable scientific gains in economics thanks to this movement.[93] The reason we did not include it on our list, though it did permeate certain aspects of our discussion (see Section 8), is that this push has taken place and is relatively far along already: the Open Science Framework's database has received 18,000 preregistrations since its launch in 2012, with the number roughly doubling every year, and more than 120 journals in various fields now offer registered reports. As such, while we strongly recommend pre-registering experiments (possibly augmented with machine learning, as recently suggested by Ludwig et al. (2019)), we feel that its inclusion in our wish list would add less value than other, less-discussed items.

Our second addition emphasizes the need to work on issues of first order import, and disseminate results from economic experiments more broadly. Academic researchers, responding to career incentives that are almost exclusively tied to publications in peer-reviewed economics journals, typically spend little time and effort communicating their findings to a wider audience. While the profession has made progress towards experiments that produce relevant, generalizable and scalable results, researchers are typically not rewarded for getting involved with the actual larger-scale implementation of their results. As a result, even the most important new scientific findings with direct practical relevance often take long to reach policymakers or practitioners, and when they do, they are often misrepresented. To change this practice, we urge researchers to make their results more available

[93]The practice of requiring detailed pre-analysis plans for all empirical work has not been unanimously endorsed by all in the profession. For instance, Coffman and Niederle (2015) warn that pre-analysis plans may discourage the use of novel research designs. Instead, Heckman and Singer (2017, p.299) recommend the practice of abduction, whereby "[t]he successful abductor immerses himself in the data and the conceptual issues underlying its generation and its interpretation, and reports the results of this immersion to the reader. It is a public process where evidence, provisional models, and methods are revealed and scrutinized."

by creating informative press briefings, posting summaries and non-gated versions of their published papers online, exploiting the opportunities offered by social media, and establishing contact with policymakers and practitioners interested in applying their findings in practice. Furthermore, we see great potential gains in engaging with the scientific community more broadly, both by working more closely with researchers from other social science disciplines, and by following the methodological discussions in other experimental sciences such as biostatistics or neuroscience.

We end our wish list with a call to the experimental economics community to continue engaging in the discussion to improve our field. While Samuelson's famous quip that doing methodological research is akin to doing calisthenics remains true today, we hope that our work provides a useful starting point for those new to this discussion. We encourage researchers who have been actively shaping this debate to create their own wish lists, or to share with us items that we have missed in ours, since choosing one's favorite methodological points is often fraught with err and oversight, and we trust that our list contains both.

References

Abadie, Alberto. 2018. Statistical non-significance in empirical economics. NBER Working Paper No. 24403.

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge. 2017. When should you adjust standard errors for clustering? NBER Working Paper No. 24003.

Acemoglu, Daron. 2010. Theory, general equilibrium, and political economy in development economics. *The Journal of Economic Perspectives*, 23(4):17–32.

Acland, Dan and Matthew R. Levy. 2015. Naiveté, projection bias, and habit formation in gym attendance. *Management Science*, 61(1):146–60.

Al-Ubaydli, Omar and John A. List. 2013. On the generalizability of experimental results in economics: with a response to Camerer. NBER Working Paper No. 19666.

Al-Ubaydli, Omar and John A. List. 2015. Do natural field experiments afford researchers more or less control than laboratory experiments? *American Economic Review*, 105(5):462–6.

Al-Ubaydli, Omar and John A. List. 2019. How natural field experiments have enhanced our understanding of unemployment. *Nature Human Behaviour*, 3(1):33–9.

Al-Ubaydli, Omar, John A. List, Danielle LoRe, and Dana Suskind. 2017a. Scaling for economists: Lessons from the non-adherence problem in the medical literature. *Journal of Economic Perspectives*, 31(4):125–44.

Al-Ubaydli, Omar, John A. List, and Dana Suskind. 2017b. The science of using science. *In preparation for the International Economic Review*.

Al-Ubaydli, Omar, John A. List, and Dana Suskind. 2017c. What can we learn from experiments? Understanding the threats to the scalability of experimental results. *American Economic Review*, 107(5):282–86.

Allcott, Hunt and Judd B. Kessler. 2019. The welfare effect of nudges: A case study of en-

ergy use social comparisons. *American Economic Journal: Applied Economics*, 11(1):236–76.

Allcott, Hunt and Todd Rogers. 2014. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104(10):3003–37.

Allcott, Hunt and Dmitry Taubinsky. 2015. Evaluating behaviorally-motivated policy: Experimental evidence from the lightbulb market. *American Economic Review*, 105(8):2501–38.

Alpizar, Francisco, Fredrik Carlsson, and Olof Johansson-Stenman. 2008. Does context matter more for hypothetical than for actual contributions? Evidence from a natural field experiment. *Experimental Economics*, 11(3):299–314.

Amrhein, Valentin and Sander Greenland. 2018. Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, 2(1):4.

Anderson, Laurie M, Toby A Quinn, Karen Glanz, Gilbert Ramirez, Leila C Kahwati, Donna B Johnson, Leigh Ramsey Buchanan, W Roodly Archer, Sajal Chattopadhyay, Geetika P Kalra, David L Katz, and Task Force on Community Preventive Services. 2009. The effectiveness of worksite nutrition and physical activity interventions for controlling employee overweight and obesity: A systematic review. *American Journal of Preventive Medicine*, 37(4):340–57.

Anderson, Michael L. 2008. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Anderson, Richard G. and Areerat Kichkha. 2017. Replication, meta-analysis, and research synthesis in economics. *American Economic Review*, 107(5):56–9.

Andrews, Isaiah and Maximilian Kasy. 2017. Identification of and correction for publication bias. NBER Working Paper No. 23298.

Angrist, Joshua. 1990. Lifetime earnings and the vietnam era draft lottery: Evidence from Social Security administrative records. *The American Economic Review*, 80(3):313–336.

Angrist, Joshua and Guido Imbens. 1994. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Angrist, Joshua, Guido Imbens, and Donald Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–55.

Arellano, Manuel and Stéphane Bonhomme. 2012. Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, 79(3):987–1020.

Arrow, Kenneth. 1973. The theory of discrimination. In Ashenfelter, O. and A. Rees, editors, *Discrimination in Labor Markets*, pp. 3–33. Princeton University Press.

Athey, Susan. 2018. The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Athey, Susan and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7353–60.

Athey, Susan and Guido Imbens. 2017a. The econometrics of randomized experiments. In Banerjee, A.V. and E. Duflo, editors, *Handbook of Economic Field Experiments*, volume 1, pp. 73–140. North-Holland.

Athey, Susan and Guido Imbens. 2017b. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Athey, Susan and Michael Luca. 2019. Economists (and Economics) in Tech Companies. *Journal of Economic Perspectives*, 33(1):209–230.

Attanasio, Orazio and Costas Meghir. 2012. Education choices in Mexico: Using a struc-

tural model and a randomized experiment to evaluate Progresa. *The Review of Economic Studies*, 79(1):37–66.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein. 2015. Behavioral hazard in health insurance. *Quarterly Journal of Economics*, 130(4):1623–67.

Balafoutas, Loukas and Matthias Sutter. 2012. Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, 335(6068):579–82.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics*, 120(3):917–62.

Banerjee, Abhijit. 2005. 'New developmenteconomics' and the challenge to theory. *Economic and Political Weekly*, 40(40):4340–4.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, Michael Walton, Abhijit Baneijee, Rukmini Baneiji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017a. From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, 31(4):73–102.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg. 2017b. Decision theoretic approaches to experiment design and external validity. *Handbook of Economic Field Experiments*, 1:141–74.

Bareinboim, Elias and Judea Pearl. 2013. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–34.

Becker, Gary S. 2010. *The economics of discrimination*. University of Chicago Press.

Bedoya Arguelles, Guadalupe, Luca Bittarello, Jonathan Martin Villars Davis, and Nikolas Karl Mittag. 2018. Distributional impact analysis: Toolkit and illustrations of impacts beyond the average treatment effect. IZA Discussion Paper No. 11863.

Behaghel, Luc, Bruno Crépon, and Thomas Le Barbanchon. 2015. Unintended effects of anonymous resumes. *American Economic Journal: Applied Economics*, 7(3):1–27.

Bellemare, Charles, Luc Bissonnette, and Sabine Kröger. 2014. Statistical power of within and between-subjects designs in economic experiments. IZA Discussion Paper No. 8583.

Bellemare, Charles, Luc Bissonnette, and Sabine Kröger. 2016. Simulating power of economic experiments: the powerBBK package. *Journal of the Economic Science Association*, 2(2):157–68.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. 2017. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10.

Benjamini, Yoav and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1).

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Benz, Matthias and Stephan Meier. 2008. Do people behave in experiments as in the field? Evidence from donations. *Experimental Economics*, 11(3):268–81.

Bernheim, B. Douglas, Andrey Fradkin, and Igor Popov. 2011. The welfare economics of default options: A theoretical and empirical analysis of 401 (k) plans. NBER Working Papers No. 17587.

Berry, James, Lucas C. Coffman, Douglas Hanley, Rania Gihleb, and Alistair J. Wilson. 2017. Assessing the rate of replication in economics. *American Economic Review*, 107(5):27–31.

Bertsimas, Dimitris, Mac Johnson, and Nathan Kallus. 2015. The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–76.

Bettis, Richard A. 2012. The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal*, 33(1):108–113.

Björkegren, Daniel and Darrell Grissen. 2018. Behavior revealed in mobile phone usage predicts loan repayment. SSRN Working Paper. http://dx.doi.org/10.2139/ssrn.2611775.

Bohnet, Iris, Alexandra van Geen, and Max Bazerman. 2016. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, 62(5):1225–34.

Bonferroni, CE. 1935. *Il calcolo delle assicurazioni su gruppi di teste.* Tipografia del Senato.

Borghans, Lex, James J. Heckman, Bart H. H. Golsteyn, and Huub Meijers. 2009. Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, 7(2-3):649–658.

Brandon, Alec, Paul J. Ferraro, John A. List, Robert D. Metcalfe, Michael K. Price, and

Florian Rundhammer. 2017. Do the effects of social nudges persist? Theory and evidence from 38 natural field experiments. NBER Working Paper No. 23277.

Bren, Linda. 2001. Frances Oldham Kelsey: FDA medical reviewer leaves her mark on history. *FDA Consumer*, 35(2):24–9.

Briesch, Amy M., Hariharan Swaminathan, Megan Welsh, and Sandra M. Chafouleas. 2014. Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, 52(1):13–35.

Briggs, Derek C. and Mark Wilson. 2007. Generalizability in item response modeling. *Journal of Educational Measurement Summer*, 44(2):131–55.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Browning, Martin and Pierre-Andre Chiappori. 1998. Efficient intra-household allocations: A general characterization and empirical tests. *Econometrica*, 66(6):1241–78.

Bruhn, Miriam and David McKenzie. 2009. In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4):200–32.

Burke, Lora E, Mindi A. Styn, Susan M Sereika, Molly B. Conroy, Lei Ye, Karen Glanz, Mary Ann Sevick, and Linda J. Ewing. 2012. Using health technology to enhance self-monitoring for weight loss: A randomized trial. *American Journal of Preventive Medicine*, 43(1):20–6.

Burke, Marshall, Lauren Falcao Bergquist, and Edward Miguel. 2014. Sell low and buy high: Arbitrage and local price effects in Kenyan markets. *The Quarterly Journal of Economics*, 134(2):785–842.

Burlig, Fiona, Louis Preonas, and Matt Woerman. 2017. Panel data and experimental design. Energy Institute at Haas Working Paper No. 277.

Butera, Luigi and John A. List. 2017. An economic approach to alleviate the crises of confidence in Science: With an application to the public goods game. NBER Working Paper No. 23335.

Button, Katherine S, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14(5):365–76.

Camerer, Colin. 2015. The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. In Fréchette, Guillaume and Andrew Schotter, editors, *Handbook of Experimental Economic Methodology*, pp. 249–95. Oxford University Press.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-hua Ho, Jürgen Huber, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–6.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric Jan Wagenmakers, and Hang Wu. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–44.

Card, David, Stefano DellaVigna, and Ulrike Malmendier. 2011. The role of theory in field experiments. *Journal of Economic Perspectives*, 25(3):39–62.

Carter, E. C. and M. E. McCullough. 2014. Publication bias and the limited strength

model of self-control: has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, 5:1–11.

Chandar, Bharat K., Ali Hortacsu, John A. List, Ian Muir, and Jeffrey M. Wooldridge. 2018. Design and analysis of cluster-randomized field experiments in panel data settings. Unpublished Manuscript.

Chandrasekhar, Arun G., Benjamin Golub, and He Yang. 2018. Signaling, shame, and silence in social learning. Technical report, National Bureau of Economic Research.

Chang, Andrew C. and Phillip Li. 2017. A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review*, 107(5):60–4.

Charness, Gary and Uri Gneezy. 2009. Incentives to exercise. *Econometrica*, 77(3):909–931.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization*, 81(1):1–8.

Chassang, Sylvain, Padró I. Miquel, Erik Snowberg, and Others. 2012. Selective trials: A principal-agent approach to randomized controlled experiments. *American Economic Review*, 102(4):1279–309.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, 106(4):855–902.

Christensen, Garret and Edward Miguel. 2018. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80.

Cleave, BL, N. Nikiforakis, and R. Slonim. 2013. Is there selection bias in laboratory experiments? The case of social and risk preferences. *Experimental Economics*, 16(3):372–82.

Clemens, Michael A. 2015. The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, 31(1):326–42.

Coffman, Lucas C. and Muriel Niederle. 2015. Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, 29(3):81–98.

Coffman, Lucas C., Muriel Niederle, and Alistair J. Wilson. 2017. A proposal to organize and promote replications. *American Economic Review*, 107(5):41–5.

Coker, Beau, Cynthia Rudin, and Gary King. 2018. A theory of statistical inference for ensuring the robustness of scientific results. Unpublished Manuscript.

Cooper, Brittany Rhoades, Brian K. Bumbarger, and Julia E. Moore. 2015. Sustaining evidence-based prevention programs: Correlates in a large-scale dissemination initiative. *Prevention Science*, 16(1):145–57.

Coville, Aidan and Eva Vivalt. 2017. How often should we believe positive results? Assessing the credibility of research findings in development economics. Working Paper.

Cox, David R and Nancy Reid. 2000. *The theory of the design of experiments*. Chapman & Hall/CRC Press.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, 128(2):531–80.

Croke, Kevin, Joan Hamory Hicks, Eric Hsu, Michael Kremer, and Edward Miguel. 2016. Do mass deworming affect child nutrition? Meta-analysis, cost effectiveness and statistical power. NBER Working Paper No. 22382.

Croson, Rachel and Uri Gneezy. 2009. Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74.

Cunha, Jesse M., De Giacomo Giorgi, and Seema Jayachandran. 2017. The price effects of cash versus in-kind transfers. *Review of Economic Studies*, 86(1):240–81.

Davis, Jonathan, Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. The economics of scale-up. NBER Working Paper No. 23925.

Davis, Jonathan M.V. and Sara B. Heller. 2017. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50.

De Long, J. Bradford and Kevin Lang. 1992. Are all economic hypotheses false? *Journal of Political Economy*, 100(6):1257–72.

Deaton, Angus and Nancy Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine*, 210:2–21.

Deck, Cary A., Enrique Fatas, and Tanya Rosenblat, editors 2015. *Replication in experimental economics*. Research in Experimental Economics. Emerald Group Publishing Limited.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, 127(1):1–56.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2016. Estimating social preferences and gift exchange at work. NBER Working Paper No. 22043.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2017. Voting to tell others. *The Review of Economic Studies*, 84(1):143–81.

DellaVigna, Stefano and Devin Pope. 2018. Predicting experimental results: Who knows what? *Journal of Political Economy*, 126(6):2410–56.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. Replication in empirical economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, 76(4):587–603.

Doucouliagos, Chris and T. D. Stanley. 2013. Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys*, 27(2):316–39.

Dreber, A., T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson. 2015. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–7.

Duflo, Esther. 2004. Scaling up and evaluation. In *Annual World Bank Conference on Development Economics 2004*, pp. 341–69.

Duflo, E., R. Glennerster, and M. Kremer. 2007. Using randomization in development economics research: A toolkit. In Schultz, TP and JA Strauss, editors, *Handbook of Development Economics*, volume 4, pp. 3895–962.

Dunn, Olive Jean. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Dupas, Pascaline. 2014. Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1):197–228.

Duvendack, Maren, Richard Palmer-Jones, and W. Robert Reed. 2017. What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review*, 107(5):46–51.

Falk, Armin and James J. Heckman. 2009. Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952):535 –38.

Feiveson, Alan H. 2002. Power by simulation. *The Stata Journal*, 2(2):107–24.

Ferraro, Paul J., Juan Jose Miranda, and Michael K. Price. 2011. The persistence of treatment effects with norm-based policy instruments: Evidence from a randomized environmental policy experiment. *The American Economic Review*, 101(3):318–22.

Ferraro, Paul J. and Michael K. Price. 2013. Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1):64–73.

Fiedler, Klaus, Florian Kutzner, and Joachim I. Krueger. 2012. The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, 7(6):661–9.

Finkelstein, Amy. 2007. The aggregate effects of health insurance: Evidence from the Introduction of Medicare. *The Quarterly Journal of Economics*, pp. 1–37.

Finkelstein, Amy and Matthew Notowidigdo. 2018. Take-up and targeting: Experimental evidence from SNAP. NBER Working Paper No. 24652.

Fisher, Ronald A. 1925. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Fisher, Ronald A. 1935. *The design of experiments*. Oliver and Boyd, London, Edinburgh.

Flory, Jeffrey A., Uri Gneezy, Kenneth L. Leonnard, and John A. List. 2018. Gender, age, and competition: a disappearing gap? *Journal of Economic Behavior and Organization*, 150(June):256–76.

Food and Drug Administration. 1997. Guidance for industry. Technical report.

Food and Drug Administration. 2017. Women's Health Research. Regulations, guidance and reports related to women's health.

Freedman, DA. 2006. Statistical models for causation: What inferential leverage do they provide? *Evaluation Review*, 30(6):691–713.

Friedman, Milton. 1953. The methodology of positive economics. In *Essays in Positive Economics*. University of Chicago Press, Chicago.

Frison, Lars and Stuart J. Pocock. 1992. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine*, 11(13):1685–704.

Fryer, Roland G., Steven D. Levitt, and John A. List. 2015. Parental incentives and early childhood achievement: A field experiment in Chicago Heights. NBER Working Paper No. 21477.

Fryer, Roland G., Steven D. Levitt, John A. List, and Anya Samek. 2017. Towards an understanding of what works in preschool education. Unpublished manuscript.

Gabaix, Xavier and Emmanuel Farhi. 2017. Optimal taxation with behavioral agents. Society for Economic Dynamics Working Paper No. 1634.

Gächter, Simon. 2010. (Dis)advantages of student subjects: What is your research question? *Behavioral and Brain Sciences*, 33(2-3):92–93.

Gelman, Andrew and John Carlin. 2014. Beyond power calculations: Assessing type S (Sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–51.

Gelman, A and J Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, New York, NY.

Gerber, Alan S., Donald P. Green, and Ron Shachar. 2003. Voting may be habit-forming: Evidence from a randomized field experiment. *American Journal of Political Science*, 47(3):540–50.

Gibson, Rajna, Carmen Tanner, and Alexander F. Wagner. 2013. Preferences for truthfulness: Heterogeneity among and within individuals. *American Economic Review*, 103(1):532–48.

Giné, Xavier, Dean Karlan, and Jonathan Zinman. 2010. Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics*, 2(4):213–35.

Glaeser, Edward L., Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review*, 106(5):114–8.

Gneezy, Uri and John A. List. 2006. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–84.

Goel, Sharad, Justin M. Rao, Ravi Shroff, and Others. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–94.

Goeree, Jacob K. and Charles A. Holt. 2001. Ten little treasures of game theory and ten intuitive contradictions. *The American Economic Review*, 91(5):1402–22.

Gosnell, Greer K., John A. List, and Robert Metcalfe. 2017. A new approach to an age-old

problem: Solving externalities by incenting workers directly. *Journal of Public Economics*, 148(April):14–31.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–50.

Greenstone, Michael and Ted Gayer. 2009. Quasi-experimental and experimental approaches to environmental economics. *Journal of Environmental Economics and Management*, 57(1):21–44.

Groh, Matthew, Nandini Krishnan, David McKenzie, and Tara Vishwanath. 2016. The impact of soft skills training on female youth employment: Evidence from a randomized experiment in Jordan. *IZA Journal of Labor and Development*, 5(1):9.

Hagger, M. S. and N. L. D. Chatzisarantis. 2016. A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4):546–73.

Hagger, Martin S., Chantelle Wood, Chris Stiff, and Nikos L. D. Chatzisarantis. 2010. Ego depletion and the strength model of self-control: a meta-analysis. *Psychological Bulletin*, 136(4):495–525.

Hallsworth, Michael, John A. List, Robert Metcalfe, and Ivo Vlaev. 2017. The behavioralist as tax collector: using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148:14–31.

Hallsworth, Michael, John A. List, Robert D. Metcalfe, and Ivo Vlaev. 2015. The making of homo honoratus: From omission to commission. NBER Working Paper No. 21210.

Hamermesh, Daniel S.. 2007. Viewpoint: Replication in economics. *Canadian Journal of Economics*, 40(3):715–33.

Hamermesh, Daniel S.. 2017. Replication in labor economics: Evidence from data and what it suggests. *American Economic Review*, 107(5):37–40.

Handel, Benjamin R.. 2013. Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, 103(7):2643–82.

Harrison, Glenn W. and John A. List. 2004. Field Experiments. *Journal of Economic Literature*, 42(4):1009–55.

Heckman, James J. 2000. Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, 115(1):45–97.

Heckman, James J. 2010. Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature*, 48:356–98.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics*, 1(1):1–46.

Heckman, James J. and Burton Singer. 2017. Abducting Economics. *American Economic Review*, 107(5):298–302.

Hedblom, Daniel, Brent R. Hickman, and John A. List. 2016. Toward an understanding of corporate social responsibility: Theory and field experimental evidence. *Unpublished manuscript*.

Heinrich, Janet. 2001. General accounting office report: GAO-01-286R Drugs withdrawn from market. Technical report.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. In search of Homo Economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review, Papers & Proceedings*, 91(2):73–78.

Henrich, J and SJ Heine. 2010. Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain*, 33(2-3):111–35.

Henrich, J, SJ Heine, and A Norenzayan. 2010a. Most people are not WEIRD. *Nature*, 466:29.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010b. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2010):61–135.

Higgins, J and SG Thompson. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–58.

Hoel, Jessica B.. 2015. Heterogeneous households: A within-subject test of asymmetric information between spouses in Kenya. *Journal of Economic Behavior & Organization*, 118:123–35.

Höffler, Jan H. 2017. Replication and economics journal policies. *American Economic Review*, 107(5):52–5.

Holm, Sture. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Horton, John, David Rand, and Richard Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425.

Hsee, Christopher K. 1996. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67(3):247–57.

Huang, Jason, David H. Reiley, and Nickolai M. Riabov. 2017. Measuring consumer sensitivity to audio advertising: A field experiment on pandora internet radio. SSRN Working Paper. http://dx.doi.org/10.2139/ssrn.3166676.

Ioannidis, John P.A. 2005. Why most published research findings are false. *PLoS Medicine*, 2(8):696–701.

Ioannidis, John P.A., Iztok Hozo, and Benjamin Djulbegovic. 2013. Optimal type I and type II error pairs when the available sample size is fixed. *Journal of Clinical Epidemiology*, 66(8):903–10.

Ioannidis, John P.A., T. D. Stanley, and Hristos Doucouliagos. 2017. The Power of Bias in Economics Research. *Economic Journal*, 127(605):F236–65.

Jackson, C Kirabo. 2010. A little now for a lot later: A look at a Texas Advanced Placement incentive program. *Journal of Human Resources*, 45(3):591–639.

Jennions, Michael D. and Anders Pape Møller. 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology*, 14(3):438–45.

Jensen, Robert. 2010. The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, 125(2):515–48.

Jimenez-Gomez, David. 2017. Nudging and phishing: A theory of behavioral welfare economics. Unpublished Manuscript.

Jimenez-Gomez, David. 2018. Hyperbolic discounting is not lack of self-control. SSRN Working Paper. http://dx.doi.org/10.2139/ssrn.3259378.

John, Leslie K., George Loewenstein, Andrea B. Troxel, Laurie Norton, Jennifer E. Fassbender, and Kevin G. Volpp. 2011. Financial incentives for extended weight loss: A randomized, controlled trial. *Journal of General Internal Medicine*, 26(6):621–626.

Karahanna, Elena, Izak Bensabat, Ravi Bapna, and Arun Rai. 2018. Opportunities and Challenges for Different Types of Online Experiments. *Management Information Systems Quarterly*, 42(04):3–10.

Karlan, Dean and John A. List. 2007. Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–93.

Kasy, Maximilian. 2016. Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(3):324–38.

Keren, Gideon. and Charles Lewis. 1993. *A handbook for data analysis in the behavioral sciences: Volume 1: Methodological issues*. Lawrence Erlbaum Associates, Inc.

Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron Gregory Adams, Jr. Reginald B. Adams, Sinan Alper, Mark Aveyard, Jordan Axt, Mayowa Babaloia, Štěpán Bahník, Mihaly Berkics, Michael Jason Bernstein, Daniel R. Berry, Olga Bialobrzeska, Konrad Bocian, Mark Brandt, Robert Busching, Huajian Cai, Fanny Cambier, Katarzyna

Cantarero, Cheryl L. Carmichael, Zeynep Cemalcilar, Jesse J. Chandler, Jen-Ho Chang, Armand Chatard, Eva CHEN, Winnee Cheong, David C. Cicero, Sharon Coen, Jennifer A. Coleman, Brian Collisson, Morgan Conway, Katherine S. Corker, Paul G. Curran, Fiery Cushman, Ilker Dalgar, William E. Davis, Maaike de Bruijn, Marieke de Vries, Thierry Devos, Canay Doğulu, Nerisa Dozo, Kristin Dukes, Yarrow Dunham, Kevin Durrheim, Matthew Easterbrook, Charles R. Ebersole, John Edlund, Alexander Scott English, Anja Eller, Carolyn Finck, Miguel-Ángel Freyre, Mike Friedman, Natalia Frankowska, Elisa Maria Galliani, Tanuka Ghoshal, Steffen Robert Giessner, Tripat Gill, Timo Gnambs, Angel Gomez, Roberto Gonzalez, Jesse Graham, Jon Grahe, Ivan Grahek, Eva Green, Kakul Hai, Matthew Haigh, Elizabeth L. Haines, Michael P. Hall, Marie E. Heffernan, Joshua A. Hicks, Petr Houdek, Marije van der Hulst, Jeffrey R. Huntsinger, Ho Phi Huynh, Hans IJzerman, Yoel Inbar, Åse Innes-Ker, William Jimenez-Leal, Melissa-Sue John, Jennifer Joy-Gaba, Roza Kamiloglu, Andreas Kappes, Heather Kappes, Serdar Karabati, Haruna Karick, Victor N. Keller, Anna Kende, Nicolas Kervyn, Goran Knezevic, Carrie Kovacs, Lacy Elise Krueger, German Kurapov, Jaime Kurtz, Daniel Lakens, Lili Lazarevic, Carmel Levitan, Jr. Neil Lewis, Samuel Lins, Esther Maassen, Angela Maitner, Winfrida Malingumu, Robyn Mallett, Satia Marotta, Jason McIntyre, Janko Medjedovic, Taciano L. Milfont, Wendy Morris, Andriy Myachykov, Sean Murphy, Koen Ilja Neijenhuijs, Anthony J. Nelson, Felix Neto, Austin Lee Nichols, Susan L. O'Donnell, Masanori Oikawa, Gabor Orosz, Malgorzata Osowiecka, Grant Packard, Rolando Pérez, Boban Petrovic, Ronaldo Pilati, Brad Pinter, Lysandra Podesta, Monique Pollmann, Anna Dalla Rosa, Abraham M. Rutchick, Patricio Saavedra M., Airi Sacco, Alexander K Saeri, Erika Salomon, Kathleen Schmidt, Felix Schönbrodt, Maciek Sekerdej, David Ricardo Sirlopu, Jeanine Skorinko, Michael A. Smith, Vanessa Smith-Castro, Agata Sobkow, Walter J. Sowden, Philipp Spachtholz, Troy G. Steiner, Jeroen Stouten, Chris NH Street, Oskar Sundfelt, Ewa Szumowska, Andrew Tang, Norbert K. Tanzer, Morgan Tear, Jordan Theri-

ault, Manuela Thomae, David Torres, Jakub Traczyk, Joshua M. Tybur, Adrienn Ujhelyi, Marcel A.L.M. van Assen, Anna van 't Veer, Alejandro Vásquez Echeverría, Leigh Ann Vaughn, Alexandra Vázquez, Diego Vega, Catherine Verniers, Mark Verschoor, Ingrid Voermans, Marek Vranka, Cheryl Welch, Aaron Wichman, Lisa A. Williams, Julie A. Woodzicka, Marta Katarzyna Wronska, Liane Young, John M. Zelenski, and Brian A. Nosek. 2018. Many labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, 1(4):443–90.

Koudstaal, Martin, Randolph Sloof, and Mirjam Van Praag. 2015. Risk, uncertainty, and entrepreneurship: Evidence from a lab-in-the-field experiment. *Management Science*, 62(10):2897–915.

Kowalski, Amanda. 2018. How to examine external validity within an experiment. NBER Working Paper No. 24834.

Lakens, Daniel, Federico G. Adolfi, Casper Albers, Farid Anvari, Matthew AJ Apps, Shlomo Engelson Argamon, Marcel A.L.M. van Assen, Thom Baguley, Raymond Becker, Stephen D. Benning, Daniel E. Bradford, Erin Michelle Buchanan, Aaron Caldwell, Ben van Calster, Rickard Carlsson, Sau-Chin Chen, Bryan Chung, Lincoln Colling, Gary Collins, Zander Crook, Emily S. Cross, Sameera Daniels, Henrik Danielsson, Lisa DeBruine, Daniel Dunleavy, Brian D. Earp, Jason D. Ferrell, James G. Field, Nick Fox, Amanda Friesen, Caio Gomes, James A. Grange, Andrew Grieve, Robert Guggenberger, Anne-Laura Van Harmelen, Fred Hasselman, Kevin D. Hochard, Mark Romeo Hoffarth, Nicholas Paul Holmes, Michael Ingre, Peder Isager, Hanna Isotalus, Christer Johansson, Konrad Juszczyk, David Kenny, Ahmed Abdelrahim Khalil, Barbara Konat, Junpeng Lao, Erik Gahner Larsen, Gerine M.A. Lodder, Jiri Lukavsky, Christopher Madan, David Manheim, Monica Gonzalez-Marquez, Stephen R Martin, Andrea E. Martin, Deborah Mayo, Randy J. McCarthy, Kevin McConway, Colin McFarland, Gustav Nilsonne, Amanda QX Nio, Cilene Lino de Oliveira, Sam Parsons, Gerit Pfuhl, Kimberly Quinn, John Sakon,

Selahattin Adil Saribay, Iris Schneider, Manojkumar Selvaraju, Zsuzsika Sjoerds, Samuel Smith, Tim Smits, Jeffrey R. Spies, Vishnu Sreekumar, Crystal Steltenpohl, Neil Stenhouse, Wojciech Świątkowski, Miguel A. Vadillo, Matt Williams, Donald Williams, Jean-Jacques Orban de Xivry, Tal Yarkoni, Ignazio Ziano, and Rolf Zwaan. 2018. Justify your alpha: A response to "redefine statistical significance". *Nature Human Behaviour*, 2(3):168–71.

Lambdin, Charles and Victoria A. Shaffer. 2009. Are within-subjects designs transparent? *Judgment and Decision Making*, 4(7):554–66.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp. 2010. Is a donor in hand better than two in the bush? Evidence from a natural field experiment. *The American Economic Review*, 100(3):958–83.

Lee, Soohyung and Azeem M. Shaikh. 2014. Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of PROGRESA on school enrollment. *Journal of Applied Econometrics*, 29(4):612–26.

Lehmann, E. L. 1993. The Fisher, Neyman-Pearson Theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424):1242–1249.

Lenz, Widukind. 1988. A short history of thalidomide embryopathy. *Teratology*, 38(3):203–15.

Levitt, Steven D. and John A. List. 2007. What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives*, 21(2):153–74.

Levitt, Steven D. and John A. List. 2009. Field experiments in economics: the past, the present, and the future. *European Economic Review*, 53(1):1–18.

Levitt, Steven D and John A List. 2011. Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. *American Economic Journal: Applied Economics*, 3(1):224–38.

Levitt, Steven D., John A. List, and Sally Sadoff. 2009. Checkmate: exploring backward induction among chess players. *American Economic Review*, 101(2):975–90.

Levitt, Steven D., John A. List, and Sally Sadoff. 2016. The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment. NBER Working Paper No. 22107.

List, John A.. 2004a. Testing neoclassical competitive theory in multilateral decentralized markets. *Journal of Political Economy*, 112(5):1131–56.

List, John A. 2004b. The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics*, 119(1):49–89.

List, John A. 2006a. Field Experiments: A bridge between lab and naturally occurring data. *The BE Journal of Economic Analysis & Policy*, 5(2).

List, John A. 2006b. The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy*, 114(1):1–37.

List, John A. 2008. Informed consent in social science. *Science*, 322(5902):672.

List, John A. 2011. Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives*, 25(3):3–16.

List, John A., Fatemah Momeni, and Yves Zenou. 2019a. Are estimates of early education programs too pessimistic? Evidence from a large-scale field experiment that causally measures neighbor effects. Unpublished Manuscript.

List, John A., Sally Sadoff, and Mathis Wagner. 2011. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14(4):439–57.

List, John A., Azeem M. Shaikh, and Yang Xu. 2019b. Multiple hypothesis testing in experimental economics. *Experimental Economics*, pp. 1–21.

Ljungqvist, Lars. 2008. Lucas critique. In Durlauf, Steven N and Lawrence E Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke.

Loken, Eric and Andrew Gelman. 2017. Measurement error and the replication crisis. *Science*, 355(6325):584–5.

Low, Hamish and Costas Meghir. 2017. The use of structural models in econometrics. *Journal of Economic Perspectives*, 31(2):33–58.

Lucas, Robert E. 1976. Econometric policy evaluations: A critique. In *Carnegie-Rochester Conference Series on Public Policy*, volume 1, pp. 19–46.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess. 2019. Augmenting Pre-Analysis Plans with Machine Learning. *AEA Papers and Proceedings*, 109:71–76.

Luedicke, Joerg. 2013. POWERSIM: Stata module for simulation-based power analysis for linear and generalized linear models. *Statistical Software Components*.

Maniadis, Zacharias and Fabio Tufano. 2017. The research reproducibility crisis and economics of science. *The Economic Journal*, 127(605):F200–8.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2014. One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review*, 104(1):277–90.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2015. How to make experimental economics research more reproducible: Lessons from other disciplines and a new proposal. In *Replication in experimental economics*, pp. 215–230. Emerald Group Publishing Limited.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2017. To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *Economic Journal*, 127(605):209–35.

McCloskey, Donald N. 1985. The loss function has been mislaid: The rhetoric of significance tests. *The American Economic Review*, 75(2):201–5.

McCullough, B. D and H. D Vinod. 2003. Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, 93(3):873–92.

McKenzie, David. 2011. Power calculations 101: Dealing with incomplete take-up. The World Bank. Development Impact Blog. Accessed on 01-02-2019.

McKenzie, David. 2012. Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99(2):210–21.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. 2019. Abandon statistical significance. *The American Statistician*, 73(1):234–45.

Meier, Stephan. 2007. Do subsidies increase charitable giving in the long run? Matching donations in a field experiment. *Journal of the European Economic Association*, 5(6):1203–22.

Miguel, E. and M. Kremer. 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217.

Milkman, Katherine L., Julia A. Minson, and Kevin G. M. Volpp. 2013. Holding the hunger games hostage at the gym: An evaluation of temptation bundling. *Management Science*, 60(2):283–99.

Mill, John Stuart. 1836. On the definition of political economy and the method of investigation proper to it. In *Collected Works of John Stuart Mill*, pp. 120–64. University of Toronto Press, Toronto.

Miller, Rupert G. 1981. *Simultaneous statistical inference*. Springer Series in Statistics. Springer New York.

Moonesinghe, Ramal, Muin J. Khoury, and A. Cecile J. W. Janssens. 2007. Most published research findings are false - But a little replication goes a long way. *PLoS Medicine*, 4(2):218–21.

Moore, Ryan T. 2012. Multivariate continuous blocking to improve political science experiments. *Political Analysis*, 20(4):460–79.

Morgan, Kari Lock and Donald B. Rubin. 2012. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–82.

Mueller-Langer, Frank, Benedikt Fecher, Dietmar Harhoff, and Gert G. Wagner. 2019. Replication studies in economics—How many and which papers are chosen for replication, and why? *Research Policy*, 48(1):62–83.

Mullainathan, Sendhil and Jann Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie, Uri Simonsohn, and Eric-Jan Wagenmakers. 2017. A manifesto for reproducible science. *Nature Publishing Group*, 1:1–9.

Muralidharan, Karthik and Paul Niehaus. 2017. Experimentation at scale. *Journal of Economic Perspectives*, 31(4):103–24.

Muralidharan, Karthik and Venkatesh Sundararaman. 2015. The aggregate effect of school choice: Evidence from a two-stage experiment in India. *The Quarterly Journal of Economics*, 130(3):1011–66.

Narita, Yusuke. 2018. Toward an ethical experiment. SSRN Working Paper. https://www.ssrn.com/abstract=3094905.

Nevo, Aviv and Michael D. Whinston. 2010. Taking the dogma out of econometrics: Structural modeling and credible inference. *Journal of Economic Perspectives*, 24(2):69–82.

Neyman, J and E. S Pearson. 1933. On the problem of the most efficienttests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337.

Niederle, Muriel, Carmit Segal, and Lise Vesterlund. 2013. How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science*, 59(1):1–16.

Niederle, Muriel and Lise Vesterlund. 2007. Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–101.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):201708274.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):4716.

Paluck, E.L and E Shafir. 2017. The psychology of construal in the design of field experiments. In *Handbook of Economic Field Experiments*, volume 1, pp. 245–68.

Peters, Jörg, Jörg Langbein, and Gareth Roberts. 2018. Generalization in the tropics-development policy, randomized controlled trials, and external validity. *World Bank Research Observer*, 33(1):34–64.

Phelps, Edmund S. 1972. The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–61.

Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–302.

Robinson, Joan. 1977. What are the questions? *Journal of Economic Literature*, 15:1318–39.

Rodriguez-Planas, N. 2012. Longer-term impacts of mentoring, educational services, and learning incentives: Evidence from a randomized trial in the United States. *American Economic Journal: Applied Economics*, 4(4):121–39.

Rogers, Todd and Erin Lynn Frey. 2016. Changing behavior beyond the here and now. In *Blackwell Handbook of Judgment and Decision Making*, pp. 726–48. 1 edition.

Romano, Joseph and Michael Wolf. 2010. Balanced control of generalized error rates. *Annals of Statistics*, 38(1):598–633.

Romano, Joseph P. and Michael Wolf. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–82.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Royer, Heather, Mark Stehr, and Justin Sydnor. 2015. Incentives, Commitments and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company. *American Economic Journal: Applied Economics*, 7(3):51–84.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Samii, Cyrus and Peter M. Aronow. 2012. On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters*, 82(2):365–70.

Samuelson, Paul A and William D. Nordhaus. 1985. *Economics*. McGraw Hill, New York, 12 edition.

Seidel, Joseph and Yang Xu. 2016. MHTEXP: Stata module to perform multiple hypothesis testing correction procedure. *Statistical Software Components*.

Senn, S. 2013. Seven myths of randomisation in clinical trials. *Statistics in medicine*.

Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Shang, Jen and Rachel Croson. 2009. A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*, 119(540):1422–39.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–66.

Slonim, R, C Wang, E Garbarino, and D Merrett. 2013. Opting-in: Participation bias in economic experiments. *Journal of Economic Behavior and Organization*, 90(June):43–70.

Smaldino, Paul E. and Richard McElreath. 2016. The natural selection of bad science. *Royal Society Open Science*, 3(9):160384.

Smith, Vernon L.. 1962. An experimental study of competitive market behavior. *Journal of Political Economy*, 70(3):322–3.

Smith, Vicki L., Saul M. Kassin, and Phoebe C. Ellsworth. 1989. Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology*, 74(2):356–9.

Spiegler, Ran. 2014. On the equilibrium effects of nudging. *Journal of Legal Studies*, 44(2):389–416.

Sterck, Olivier. 2018. On the economic importance of the determinants of long-term growth. *CSAE Working Paper Series*, 2018-20.

Sterne, Jonathan A.C. and George Davey Smith. 2001. Sifting the evidence—what's wrong with significance tests? *BMJ*, 322(7280):226.

Stoop, Jan, Charles N Noussair, and Daan Van Soest. 2012. From the lab to the field: Cooperation among fishermen. *Journal of Political Economy*, 120(6):1027–56.

Sukhtankar, Sandip. 2017. Replications in development economics. *American Economic Review*, 107(5):32–6.

Todd, Petra E. and Kenneth I. Wolpin. 2006. Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American Economic Review*, 96(5):1384–417.

Travers, Justin, Suzanne Marsh, Mathew Williams, Mark Weatherall, Brent Caldwell, Philippa Shirtcliffe, Sarah Aldington, and Richard Beasley. 2007. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax*, 62(3):219–23.

Vivalt, Eva. 2017. How much can we generalize from impact evaluations? Unpublished Manuscript.

Volpp, Kevin, Andrea B. Troxel, Mark V. Pauly, Henry A. Glick, Andrea Puig, David A. Asch, Robert Galvin, Jingsan Zhu, Fei Wan, Jill DeGuzman, Elizabeth Corbett, Janet Weiner, and Janet Audrain-McGovern. 2009. A randomized, controlled trial of financial incentives for smoking cessation. *The New England Journal of Medicine*, 360(7):699–709.

Volpp, Kevin G., Andrea Gurmankin Levy, David A. Asch, Jesse A. Berlin, John J. Murphy, Angela Gomez, Harold Sox, Jingsan Zhu, and Caryn Lerman. 2006. A randomized controlled trial of financial incentives for smoking cessation. *Cancer Epidemiology, Biomarkers & Prevention*, 15(1):12–8.

Volpp, Kevin G, Leslie K John, Andrea B Troxel, Laurie Norton, Jennifer Fassbender, and George Loewenstein. 2008. Financial incentive – based approaches for weight loss: A randomized trial. *Journal of the American Medical Association*, 300(22):2631–2637.

Wacholder, Sholom, Stephen Chanock, Laure El, and Nathaniel Rothman. 2004. Assessing the probability that a positive report is. *Cancer Research*, 96(6):434–42.

Wager, Stefan and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–42.

Walton, Gregory M. and Geoffrey L. Cohen. 2011. A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, 331(6023):1447–51.

Wasserstein, Ronald L. and Nicole A. Lazar. 2016. The ASA's statement on p -Values: Context, process, and purpose. *The American Statistician*, 70(2):129–33.

Wilhelm, Daniel, Sokbae Lee, and Pedro Carneiro. 2017. Optimal data collection for randomized control trials. Cemmap Working Paper No. CWP45/17.

Young, Alwyn. 2016. Improved, nearly exact, statistical inference with robust and clus-

tered covariance matrices using effective degrees of freedom corrections. *Unpublished manuscript, London: London School of Economics and Political Science.*

Young, Alwyn. 2017. Consistency without inference: Instrumental variables in practical application. *Unpublished manuscript, London: London School of Economics and Political Science.*

Young, Alwyn. 2019. Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–98.

Zhang, L and Andreas Ortmann. 2013. Exploring the meaning of significance in experimental economics. UNSW Australian School of Business Discussion Paper No. 2013-32.

Zhe Jin, Ginger, Andrew Kato, and John A. List. 2010. That's news to me! Information revelation in professional certification markets. *Economic Inquiry*, 48(1):104–22.

Ziliak, Stephen T. and Deirdre N. McCloskey. 2004. Size matters: The standard error of regressions in the American Economic Review. *The Journal of Socio-Economics*, 33(5):527–46.

Appendix

Generalizability Threat III: noncompliance case

We consider here the more general case of Threat III to generalizability (see Section 1), for the case of noncompliance. The issue of selective noncompliance can be exacerbated by non-random selection into the experiments, in which case p_i is correlated with (z_i, d_i).[94] In that case, there is a problem of generalizability when rolling the program to the entire population.[95] If the researchers are interested in the effect of the program $\lambda^* = \mathbb{E}[y_{i1} - y_{i0}|\omega^{FFE}, d(z_i = 1) = 1, d(z_i = 0) = 0]$, we have that

$$\lambda^* = \mathbb{P}[p_i = 1] \cdot \text{LATE}_{p=1} + \mathbb{P}[p_i = 0] \cdot \text{LATE}_{p=0},$$

where $\text{LATE}_{p=1}$ is defined as in Equation 1, and $\text{LATE}_{p=0}$ is defined analogously for those with $p_i = 0$.[96] We can calculate the bias as we did in Equation 2 for the case of non-compliance, which is given by $\mathbb{P}[p_i = 0] \cdot (\text{LATE}_{p=1} - \text{LATE}_{p=0})$.[97] If the value of $\text{LATE}_{p=0}$ is very different from $\text{LATE}_{p=1}$, then the estimate from the FFE is not generalizable because, for most FFE, $\mathbb{P}[p_i = 0]$ is much larger than $\mathbb{P}[p_i = 1]$. However, the estimate from FFE will be generalizable when $\text{LATE}_{p=1} \approx \text{LATE}_{p=0}$, and this can happen if p is independent of (z_i, d_i):

$$p_i \perp\!\!\!\perp (z_i, d_i)|x_i \qquad \text{(Compliance Independence Condition)}.$$

[94] For example, if the stakes involved may affect both the selection decision into the experiment p_i, and then the subsequent behavior in response to the treatment d_i, or if those who select into the experiment are more likely to comply with their assigned treatment, or the opposite, those who select into the experiment are more likely to choose a particular d no matter what z_i is.

[95] See also Section 12 for issues on scalability.

[96] That is, $\text{LATE}_{p=0} = \mathbb{E}[y_{i1} - y_{i0}|\omega^{FFE}, d_i(z_i = 1) = 1, d_i(z_i = 0) = 0, p_i = 0]$.

[97] Because $\text{LATE}_{p=1} - \lambda^* = (1 - \mathbb{P}[p_i = 1]) \cdot \text{LATE}_{p=1} - \mathbb{P}[p_i = 0] \cdot \text{LATE}_{p=0}$.

Generalizability framework from Al-Ubaydli and List (2013)

We present a very similar version of the framework in Al-Ubaydli and List (2013), which was based on Heckman (2000)].

We say D has **local generalizability** if

$$\forall (x, x', z) \in D, \exists \epsilon > 0 : B_\epsilon(x, x', z) \subset D \cup \Delta(R).$$

Note that if D is an open set, then D has local generalizability, even if D has zero generalizability.[98] We say that D has **global generalizability (of size M)** if $D^M \subset D \cup \Delta(R)$, where

$$D^\epsilon = \{(x, x', z) : \exists (\bar{x}, \bar{x}', \bar{z}) \in D \text{ with } (x, x', z) \in B_\epsilon(\bar{x}, \bar{x}', \bar{z})\}.$$

Note that global generalizability of size $M > 0$ implies local generalizability. Moreover, if D is finite, local generalizability implies global generalizability for some $M > 0$.[99]

Attenuation bias over time

Let $\tau_t = \mathbb{E}[y_{it} - y_{it_0} | z_{i0} = 1] - \mathbb{E}[y_{it} - y_{it_0} | z_{i0} = 0]$. Since z_{i0} is random, then we have that $\mathbb{E}[y_{it_0} | z_{i0} = 1] = \mathbb{E}[y_{it_0} | z_{i0} = 0]$, and hence

$$\tau_t = \mathbb{E}[y_{it} | z_{i0} = 1] - \mathbb{E}[y_{it} | z_{i0} = 0].$$

We assume that each individual changes treatments with a constant probability that is positive if the new treatment has a higher outcome, and zero otherwise. For each individual, this generates a Markov Chain, and the stationary distribution π_i is such that there

[98] This is because if D is open, then around each point $(x, x', z) \in D$ we can always find a small enough open ball around (x, x', z) that is contained in D.

[99] In particular, for $M = \min\{\epsilon : B_\epsilon(x, x', z) \subset D \cup \Delta(R) \forall (x, x', z) \in D\}$.

is a mass 1 of probability at the treatment with the highest outcome (it is an absorbing state) for individual i. Therefore, as $t \to \infty$, each individual converges to the treatment with the highest outcome. Again, as a result of z_{i0} being random, the stationary distribution π_i is the same in expectation for $z_{i0} = 1$ and $z_{i0} = 0$, and therefore:

$$\lim_{t\to\infty} \tau_t = \lim_{t\to\infty} \mathbb{E}[y_{it}|z_{i0} = 1] - \mathbb{E}[y_{it}|z_{i0} = 0] = \mathbb{E}[y \cdot \pi_i|z_{i0} = 1] - \mathbb{E}[y \cdot \pi_i|z_{i0} = 0] = 0.$$