Site Selection and External Validity in Observational and Experimental Settings: How a "risky" RCT may underperform OLS [PRELIMINARY AND INCOMPLETE]

Christian Ahlin*

July 2023

Abstract

Randomized, controlled trials (RCTs) are often thought to generate the single most credible form of "causal" evidence. We consider a setting in which the population is divided into sites, and an experimenter must find a willing site to implement and test a potentially risky treatment. We compare the potential biases in estimating the population average treatment effect (ATE) using an experimental approach and an analogous observational approach. A canonical site selection bias – based on a site's forecast of treatment effect – may plague both approaches. If so, it is a problem of internal validity in the observational case but external validity in the experimental case. We model selection in both the observational and the experimental contexts, and provide conditions under which the ATE is estimated with greater bias using the experimental rather than the observational approach. We conclude that in the context of a site-based, risky treatment, the evidence from even a much-replicated RCT need not be more broadly informative than that of an observational study. Skeptical sites are on standard econometric ground treating such experimental evidence as "non-causal".

*Department of Economics, Michigan State University; +1 517 3558306; ahlinc@msu.edu.

1 Introduction

Consider bringing empirical evidence to bear on an open policy question in the field of microfinance, such as whether group lending generally raises repayment rates, whether regular or backloaded repayment schedules lead to more efficient and sustainable outcomes, or whether equity contracts tend to outperform debt ones. One option may be simply to compare outcomes of microfinance institutions (MFIs) that use different approaches, for example group- and non-group based lenders. Many would consider this option significantly inferior to an experimental approach involving partnering with an MFI to randomly assign the two practices to two subsets of the MFIs' clientele, and comparing outcomes. Rated even better would be accumulated evidence from this experiment replicated on a number of MFIs.

More generally, experimental studies ("RCTs") are often thought to generate the single most credible form of empirical evidence on causal impact,[1] likely due to their ability to produce an unbiased estimate of the average treatment effect ("ATE") under minimal assumptions.[2]

This paper argues that in certain natural settings, there is no obvious reason to expect the experimental approach to produce a less biased estimate of the population ATE than a simple observational approach. The settings we focus on involve subsets of the population called "sites" – states, NGOs, firms, banks, etc. – that are responsible for determining treatment. The average treatment effect may vary across sites, and each site has a benevolent decision-maker that decides whether to undertake treatment. The decision-maker's objective coincides with the outcome of interest to the researcher. Crucially, the treatment is risky: it has a real chance of appreciably worsening outcomes at the site, as in each of the microfinance examples above.

A simple observational approach in such a setting would compare outcomes at sites that adopted the treatment and sites that did not. As is well-known, selection bias may plague

[1] See Imbens (2010), for example.

[2] See Athey and Imbens (2017) and Deaton and Cartwright (2018), for example.

the resulting estimate. The potential bias can be decomposed into "Bias–1", the difference in treatment effect at sites selecting treatment versus the average site, and "Bias–2", the difference in untreated outcomes between sites selecting into and out of treatment. The potential for either bias is often considered sufficient to relegate evidence from this approach to "descriptive" (or "non-causal") status.

The RCT approach involves a research team partnering with a site to experimentally test the treatment. Randomization would eliminate (in expectation) the differences in untreated outcomes between treated and untreated in the treatment site, eliminating an analog of "Bias–2". But a potential bias remains, an analog of "Bias–1": the difference in treatment effect at sites willing to select into experimental treatment and the average site. This bias does not threaten internal validity of the study, but would be classified as an external validity issue. Still, it would bias the experimental estimate of the population ATE; and this bias would remain after many replications.

"Bias–1" in both the observational and the experimental cases is the selection bias stemming from the willingness of sites to select into the treatment. In one case selection is into autonomous treatment while in the other it is into experimental treatment; but a natural assumption would be that optimizing site-level decision-makers take their own forecasts of the treatment effect into account in both cases, and especially when the treatment carries significant risk. Thus, presumptively both the observational approach and the experimental approach suffer from a similar, canonical selection bias with respect to the population ATE. If so, the superiority of the experimental approach for estimating the ATE is partly a mirage: it does not eliminate a potentially critical bias, but instead shifts it from the realm of internal validity to external validity where it receives less scrutiny.

The experimental approach does not merely shift the analog of Bias–2, but eliminates it with randomization. This may be considered reason to prefer the experimental approach – it gets rid of one potential bias, if not two. But this logic does not hold up. We explore implications for bias in a Roy model of selection into treatment in both autonomous and

experimental settings. Many considerations could enter such a model, but we aim for a simple framework that highlights key issues. Under some conditions – e.g. limited patience of the site manager, limited trust in the experimenting team or the experimental process, knowledge free-riding problems – the selection problem is indeed the same in both the autonomous and the experimental settings. In this case, the observational approach can produce a less biased estimate of the ATE than the experimental approach, depending on the distribution of heterogeneity across sites. This is because Bias–1 and Bias–2 may counteract, so that the sum of both biases may be closer to zero than Bias–1 by itself.

In sum, a "risky" experiment – even a much-replicated one – may produce no better evidence on the population ATE than an observational approach. It may be argued that the parameter of interest is the site-specific ATE, not the population ATE as assumed above. This is undoubtedly the case in some contexts (Deaton and Cartwright, 2018). But as a broad statement, it belies the use and popularity of the experimental approach, whose main general interest arguably stems from the ability to create general insights into what works or how things work more broadly, across contexts (e.g., see Imbens, 2010, p. 417). Yet in the aspiration to produce scientific evidence of policy relevance beyond its context, a risky RCT may be unable to shake one of the potential biases considered canonical in observational work.

The experimental approach can in principle produce an unbiased estimate of the ATE among the willing population, i.e. among sites that are willing to opt into the experiment. We believe experimental studies should make clear that the "causal" interpretation of their results applies only to such sites. In some cases, this may reasonably be assumed to include virtually all sites. But in other cases, particularly with a "risky" experiment, skeptical, unwilling sites may well abound, and a presumption that by virtue of randomization RCTs deliver more credible evidence for them than an analogous OLS study appears unfounded.

Related, this perspective provides a caveat to the finding that the randomized approach is especially suited for convincing an adversarial audience (Banerjee et al., 2020). Their logic

applies to the site-specific ATE, but given the logic of site selection, skeptical site decision-makers – ones that would not have agreed to the experiment – are on standard econometric ground in downplaying the implications of RCT evidence for themselves.[3] It also casts doubt on the idea that RCTs are most useful in testing out innovative modes of operation or techniques (e.g. Morduch, 2020). Unless the innovations are likely to be trivial or positive in impact – i.e. unless the experiment is non-"risky" – current empirical standards entail the worry that, from the perspective of general advice beyond the implementing site, RCTs generate biased results in a similar class as correlational observational studies.

The paper is organized as follows. Section 2 discusses related literature. Section 3 sets out the statistical model and provides a preliminary comparison of the biases in the observational and experimental cases. Section 4 models selection into treatment in both cases, and combines results with the statistical model to provide conditions under which each approach is more biased. Section 6 discusses implications and concludes.

2 Relation to the Literature

This paper focuses on the interaction of external and internal validity across observational and experimental studies in certain settings where site selection is a concern.

External validity is a common topic in the methodological literature discussing experiments. Multiple authors have made the criticism that far more attention is paid to internal validity than external validity, that is, to whether an estimate is valid within the study context than to whether and how it informs the world outside the realm of the study (e.g. Deaton and Cartwright, 2018), even though both questions are critical for policy (e.g. Manski, 2013). Advocates of experiments respond that external validity challenges are not unique to experiments (e.g. Banerjee and Duflo, 2009) and are likely to be practically solveable by replication across diverse sites.

[3]Banerjee et al. (2017) provide a general argument against the possibility for generalized conclusions, based on the necessity to defend against a large set of priors regarding the relationship between sites. Our argument differs in focusing on canonical selection issues.

One of the threats to external validity is selection into the experiment, as Heckman (1992, 2020) seminally discussed. While Heckman and others have focused primarily on individual-level selection issues, he and Hotz (1992) also discussed site selection. In a more recent paper, Allcott (2015) argues that bias with respect to the population parameter can arise due to systematic selection of sites into experimental participation, and make the point that the selection-on-observables assumption required to eliminate this bias is formally similar to the assumption required for unbiasedness of observational impact estimates; see also Muller (2015). Similar arguments around site selection bias can be found in Heckman and Vytlacil (2007), Banerjee and Duflo (2009), Pritchett and Sandefur (2013), Fischer and Karlan (2015), Banerjee et al. (2017), Deaton and Cartwright (2018), and Czibor et al. (2019), who say that "researchers must explicitly consider selection into the experiment, in order to derive general conclusions".

These papers make similar broad points but do not delve more deeply into the nature of the site selection bias and its relationship to well-known biases in observational settings. Our paper aims to explore and expand the reasoning of these contributions regarding issues researchers must consider when generalizing from experiments, in a richer theoretical framework. Our focus is narrower: not experiments in general, but a certain class of experiments, namely "risky", site-based experiments that involve adopting a new treatment. It is with these types of experiments that canonical selection – directly connected to effectiveness of the treatment – at the site level is hardest to assume away. Thus, while the literature tends to focus on incidental reasons for site selection into experimentation,[4] we differ in focusing on a canonical selection bias. To our knowledge, this work is the first to highlight a fundamental similarity between potential site-level biases of both observational and experimental approaches: that with rational actors, site selection into autonomous risky treatment and site selection into experimental risky treatment generally involve similar effect-forecasting cal-

[4]Examples are selection based on willingness to experiment in general (implicitly, Banerjee and Duflo, 2009), efficiency of operation to evaluate and run an experiment (Heckman and Vytlacil, 2007), or alignment with experimenter goals (Allcott, 2015).

culations, so that the same canonical selection problem can lurk behind both observational and experimental studies.

We also add to the debate on the relative importance of internal and external validity and the relative vulnerability of different methods to problems in these areas. We embed both aspects of validity into a common model with a clearly specified parameter of interest; this allows for the juxtaposition and aggregation of biases – due to external and internal validity separately – across different types of studies with respect to the population ATE. This leads to the novel point that the experimental approach may not eliminate a fundamental bias plaguing observational studies, but rather shift it from the realm of internal validity to external validity. The implication is that it can be a costly oversimplification to lump all methods together as having similar external validity issues.[5]

We also extend the literature by adapting the Roy model to study the site selection process and understand how and when it may bias results from different methodologies. Combining the described elements leads to a novel theoretical exploration of plausible conditions under which an RCT, even a much-replicated RCT, may do no better in terms of bias than a simple observational approach to estimating a population ATE.

The adoption of Roy model logic for selection into experiments is not new, but can be found in Heckman et al. (1999), Heckman and Vytlacil (2007), and Athey and Imbens (2017) along with discussions of econometric implications and strategies. On the theoretical side, Malani (2008) develops a model of individual selection into medical trials and shows that an RCT estimate is biased upward relative to the treatment effect for the population that would select into the new treatment based on the current state of knowledge; Belot and James (2014) extend these results. This strand of the Roy model literature focuses on *individual* selection into experiments, while we focus on site-based selection. While key logic is similar, our approach differs in highlighting site-level decision-making and enabling quantification of internal and external validity biases, allowing for comparisons of observational

[5]Our argument does not rely on the idea that observational studies have larger scope (e.g. Dehejia, 2015); it applies even to a much-replicated RCT.

and experimental estimates on each dimension and in total. Further, the econometric strategies developed in the literature for addressing individual-level selection, e.g. estimating ITT or LATE parameters, may be much less practical to implement as solutions to site-based selection.

Hotz et al. (2005) do not focus on site selection per se, but derive conditions under which experimental results can be extrapolated to other populations; one condition is that selection into the experimental sample is independent of potential outcomes, conditional on observables (see also Allcott, 2015, for a weaker condition). Andrews and Oster (2019) and Gechter (2022) extend this line of work with assumptions on the behavior of unobservables relative to observables. Our paper ignores mediating observables, partly for simplicity but mainly because the popularity of RCTs may be credited largely to their ability to avoid selection-on-observables assumptions and their cousins for identification. A retreat to the necessity of these kinds of assumptions would undermine the case for RCTs being categorically different from observational approaches with respect to causal identification, as others have argued.

Several studies find ways to empirically assess site selection bias or other external validity issues. Allcott's (2015) seminal study provides empirical evidence of significant site selection in several settings, and its bias of ATE estimates in one setting due apparently to substantial site selection on unobservables. He does not find improvement from an observational estimate (different from the one we study). Belot and James (2016) study selection into experiments in a school nutrition program setting, and find that while few sites opt into the experiment, there is little evidence for selection on observables or bias of estimated treatment effects. Other work includes Pritchett and Sandefur (2014, 2015), who show that a more biased in-context observational estimate may dominate several unbiased out-of-context experimental estimates in terms of RMSE (see also Ravallion, 2020); and the Bayesian meta-analyses of Meager (2019) and Vivalt (2020). While the current paper does not have empirical

evidence to add to this growing literature,[6] we view it as helping to interpret empirical results from different methods and, in particular, highlighting the kinds of experiments that may well offer little improvement over basic observational methods in generating credible general knowledge.

A general argument in the literature is that experiments are especially valuable because they allow the researcher control over the mechanism for assignment to treatment (e.g. Imbens, 2010). This paper presents a caveat: if it is necessary to find a willing site to partner with, the researcher's control over the assignment mechanism may be hampered in a similar way and for some similar reasons as with a decentralized assignment mechanism.

Overall, the paper builds on a number of well-known ideas and further develops their implications for the relative merits of experimental methodology when site selection due to high stakes is a possible concern.

3 Statistical Model

Population set $\mathcal{P}$ contains N individuals, indexed by i. Let $T_i \in \{0, 1\}$ give individual i's treatment status under the policy to be evaluated. Individual i's potential outcome as a function of treatment is $Y_i(T_i)$, also written as $Y_{i0} \equiv Y_i(0)$ and $Y_{i1} \equiv Y_i(1)$. Individual i's treatment effect is $\tau_i \equiv Y_{i1} - Y_{i0}$. The parameter of interest is assumed to be the population average treatment effect "ATE", $\overline{\tau} \equiv E(\tau_i)$.

The population is spread over a set $\mathcal{S}$ of S "sites", indexed by s, interpreted as states, locales, NGOs, firms, banks, or contexts. The set of individuals at site s is $\mathcal{P}_s$, where $\{\mathcal{P}_1, ..., \mathcal{P}_S\}$ partitions $\mathcal{P}$. Let S_i be the site in which individual i is located: $S_i = \{s \in \mathcal{S} : i \in \mathcal{P}_s\}$. For simplicity, the sites are equal in population.

Sites have two key features. First, potential outcomes may be correlated within sites. For example, NGOs or firms may operate differently and in different contexts, leading to different treatment effects of the policy across sites. Define $\overline{\tau}_s$ as the average treatment

[6] Arguably, none of the studies cited involves a clear-cut example of a "risky" experiment.

effect in site s, the "sATE": $\overline{\tau}_s \equiv E(\tau_i|S_i = s)$. The assumptions guarantee that $E_{\mathcal{S}}(\overline{\tau}_s) = \overline{\tau}$.

Second, treatment status of individuals within a site is determined by a site-level decision-maker, or "DM". For example, states determine many educational policies such as rules for hiring teachers; microfinance NGOs determine whether to offer group-based or individual-based lending; firm managers decide on compensation policy; and so on.[7]

3.1 Observational Approach

Assume that treatment status T_i and the associated outcome $Y_i(T_i)$ are observed for a random sample of individuals i in $\mathcal{P}$. Further, assume that some sites have fully implemented the policy while others have not implemented it at all. Thus, the set of sites can be partitioned into two subsets, $\{\mathcal{S}_0, \mathcal{S}_1\}$, such that the sites in $\mathcal{S}_1$ ($\mathcal{S}_0$) have (have not) implemented the policy: $T_i = 1$ if $S_i \in \mathcal{S}_1$ and $T_i = 0$ if $S_i \in \mathcal{S}_0$.

A simple observational approach compares outcomes of a random sample of treated and untreated to obtain $\hat{\tau}_{obs}$, where

$$\begin{aligned} E(\hat{\tau}_{obs}) &= E\left(Y_{i1}|T_i = 1\right) - E\left(Y_{i0}|T_i = 0\right) \\ &= E\left(Y_{i1}|S_i \in \mathcal{S}_1\right) - E\left(Y_{i0}|S_i \in \mathcal{S}_0\right) \\ &= E\left(Y_{i1}|S_i \in \mathcal{S}_1\right) - E\left(Y_{i0}|S_i \in \mathcal{S}_1\right) + E\left(Y_{i0}|S_i \in \mathcal{S}_1\right) - E\left(Y_{i0}|S_i \in \mathcal{S}_0\right) \\ &= E(\tau_i|S_i \in \mathcal{S}_1) + E\left(Y_{i0}|S_i \in \mathcal{S}_1\right) - E\left(Y_{i0}|S_i \in \mathcal{S}_0\right) \\ &= \overline{\tau} + \underbrace{E(\tau_i|S_i \in \mathcal{S}_1) - \overline{\tau}}_{Bias-1} + \underbrace{E\left(Y_{i0}|S_i \in \mathcal{S}_1\right) - E\left(Y_{i0}|S_i \in \mathcal{S}_0\right)}_{Bias-2} . \end{aligned} \tag{1}$$

The potential bias of this observational approach can be decomposed into two parts. Bias–1 arises if the treatment effect in sites that opt in tends to differ from the treatment effect in sites that opt out of the policy. This bias disappears if all sites have the same treatment effect, $\overline{\tau}_s = \overline{\tau}$, $\forall s$, or if sites' treatment effects are uncorrelated with their adoption of the

[7] We abstract from the issue of individual-level selection into treatment, and assume all individuals in a treated site are treated. If there were individual-level selection as well, by necessity the focal parameter would likely become the ITT or a LATE; similar issues to those raised in this paper would be relevant.

policy, $E(\tau_i|S_i \in \mathcal{S}_1) = E(\tau_i|S_i \in \mathcal{S}_0)$.[8] Bias–2 arises if sites that opt in and sites that opt out would tend to have different outcomes without the policy, e.g. due to differences in efficiency. The possibility for either bias is widely seen as preventing simple correlational estimates from being interpreted as "causal".[9]

3.2 Experimental Approach

Assume the researcher partners with the site-s DM to assign treatment to one random subset of the site population and deny treatment to a different random subset. The experimental, RCT approach then compares the treated and untreated subsets in site s to obtain $\hat{\tau}_{rct,s}$ where

$$\begin{aligned}
E(\hat{\tau}_{rct,s}) &= E[Y_{i1}|T_i = 1, S_i = s] - E[Y_{i0}|T_i = 0, S_i = s] \\
&= E[Y_{i1}|T_i = 1, S_i = s] - E[Y_{i0}|T_i = 1, S_i = s] \\
&\qquad + E[Y_{i0}|T_i = 1, S_i = s] - E[Y_{i0}|T_i = 0, S_i = s] \\
&= \overline{\tau}_s + E[Y_{i0}|T_i = 1, S_i = s] - E[Y_{i0}|T_i = 0, S_i = s] \qquad (2) \\
&= \overline{\tau} + \underbrace{\overline{\tau}_s - \overline{\tau}}_{Bias-1} + \underbrace{E[Y_{i0}|T_i = 1, S_i = s] - E[Y_{i0}|T_i = 0, S_i = s]}_{Bias-2} \\
&= \overline{\tau} + \underbrace{\overline{\tau}_s - \overline{\tau}}_{Bias-1} + \underbrace{0}_{Bias-2} .
\end{aligned}$$

The third and fifth equalities follow from random assignment of individuals to treatment in site s. This ability to obtain an unbiased estimate of the sATE can be considered the effectiveness of RCTs in solving internal validity issues. However, a potential bias remains, Bias–1, that if non-zero would be classified as an external validity issue because it involves how to relate a parameter that is cleanly identified within the scope of the study to the

[8] See Allcott (2015) and Heckman and Vytlacil (2007).

[9] "A major concern ... is that simple comparisons between economic agents in the various regimes are often not credible as estimates of the average effects of interest because of the potential selection bias that may result from the assignment to a particular regime being partly the result of choices by optimizing agents", Imbens (2010); "A central problem is selection, the fact that participants may be systematically different from nonparticipants", Banerjee and Duflo (2009).

population parameter of interest.

Bias–1 is zero if there is no site-specific heterogeneity in treatment effects. It is also zero in expectation if site s is randomly chosen among sites, since $E_{\mathcal{S}}(\overline{\tau}_s) = \overline{\tau}$, implying that $E_{\mathcal{S}}(\hat{\tau}_{rct,s}) = \overline{\tau}$. In this case, the experimental estimate is unbiased for the ATE, and replication across sites is useful for increasing precision of estimates of the ATE.

However, sites may have a say in whether they are experimented upon and how, and thus randomness of site selection may not hold.[10] Let $\{\mathcal{S}_1^E, \mathcal{S}_0^E\}$ be the partition of $\mathcal{S}$ into sites that are willing to engage in this experiment ($\mathcal{S}_1^E$) and those that are not ($\mathcal{S}_0^E$), and assume that site s was randomly chosen among $\mathcal{S}_1^E$. Then by reasoning analogous to the above, the RCT approach produces an estimate $\hat{\tau}_{rct}$ where

$$E(\hat{\tau}_{rct}) = \overline{\tau} + \underbrace{E(\tau_i | S_i \in \mathcal{S}_1^E) - \overline{\tau}}_{Bias-1} \ . \tag{3}$$

Thus, given voluntary site selection into the experiment, the RCT may give a biased estimate of the parameter of interest, $\overline{\tau}$. It is unbiased for the the sATE, $\overline{\tau}_s$, and perhaps for the average treatment effect of sites willing to experiment, $E(\tau_i | S_i \in \mathcal{S}_1^E)$. But moving beyond these to the population ATE, $\overline{\tau}$, is often necessary if one wants to generate broader policy implications.[11] If we are unwilling to assume that selection into the experiment is independent of the site treatment effect, then an RCT estimate is no more guaranteed to be an unbiased estimate of the population ATE than an observational estimate.

3.3 Comparing the Approaches

The "Bias–1" of equations 1 and 3 is not identical, but conceptually related. Bias–1 in the observational case is related to sites' decision to opt into the policy autonomously;

[10] It fails in systematic ways in Allcott's (2015) setting, for example.

[11] Imbens makes a similar point (2010, p.417). Deaton and Cartwright (2018) point out cases where the sATE is all researchers need to care about. But it seems clear that experimental studies are typically framed, published, and cited for their putative insights that generalize beyond the context studied; conversely, it appears rare for an RCT study to foreground inability to speak beyond the sATE or the ATE for *willing* sites with "causal" force.

Bias–1 in the experimental case is related to sites' decision to opt into the policy via an experiment. While these are not the same decisions, it would be surprising in many settings if similar calculations did not enter into both. In both cases the decision is whether to opt into treatment, and optimizing agents presumably base that decision on the expected treatment effect. That is, the framework above suggests that a standard, Roy-model selection potentially surfaces in both cases. We formally model these decisions in the next Section; for now, we discuss circumstances under which we would expect the two biases to be similar, and implications in that case.

Is Bias–1 similar in the two cases? Bias–1 is presumptively a concern in observational studies. Arguably, Bias–1 is not a concern in some kinds of experiments. For example, in some settings participation in an experiment is not subject to consent. Examples are online retail firms experimentally varying prices or web layout, and researchers submitting fabricated resumes. Site selection into the experiment is controlled by the experimenter, who can often select the site(s) at random, eliminating Bias–1.

However, many RCTs require site-level consent to participate. For some experiments, consent may reasonably be considered a foregone conclusion. One example is when the treatment is a universally valued good, e.g. unconditional cash transfers to households or budgetary support to a government agency. Another possible example is when the experimentee is not required to do anything substantively new or different, but just to allow a research team to track and measure its standard product or policy.[12]

Our focus is instead on RCTs in which the site DM consents to adopt a new product or policy variation that involves risk. Assent to such a "risky" experiment brings with it a treatment that jeopardizes – i.e. may materially worsen – outcomes that the experimentee cares about. This kind of experiment is common in development economics, for example, where researchers often attempt to convince an NGO or firm or government agency to test a

[12]Willingness to have one's product evaluated can differ across potential experimentees, however, though in a way that may differ qualitatively from selection into treatment. Some reluctance may be attributed to non-trivial operational costs of coordinating with an outside evaluator; see Heckman (1992, 2020).

new product or policy variation or management approach. For example, consider a researcher that approaches a microfinance institution ("MFI") with a proposed modification to its business model – a change in contract term or structure, in the role or operation of the microfinance group, in the monitoring and incentive mechanisms employed by the lender, in credit application evaluation, etc. Or, consider a firm that is proposed a treatment involving a change in human resource policy, advertising, or the use of outside consultants. In such cases, the experimental treatment itself may significantly raise or lower the outcomes that the site cares about, whether enterprise formation or repayment rates in the case of an MFI, or costs, revenues, reputation, and overall financial success in the case of the MFI or firm.

Given this risk, assent to a risky experiment may be far from automatic. Anecdotes abound of NGOs being unwilling to test some or all proposed variations of a product,[13] and of how difficult it can be for the researcher to get to "yes". In fact, in the case of a risky experiment, it would be surprising if this were not the case, under the assumption that bureaucrats or NGOs or firm managers are maximizing agents looking out for the well-being of their clients, bearing some responsibility for their outcomes, and possessing some working knowledge – not fully eclipsed by the outside researcher's – of what effective operation looks like in context.[14] A site-specific assessment of the risks and rewards of the proposed treatment will likely factor into the willingness of the site DM to undergo treatment via experimentation.

Arguably, some experiments should be considered risky even if the possible measured treatment effects are confined within a positive range; the overhead cost of engaging in the experiment may lower the net impacts into a range spanning zero. While experimenters often cover many direct costs of implementing the experiment, there is typically a training and management cost component borne by the site. Management costs of negotiating parameters of the experiment, estimating whether it is a propitious undertaking for the organization,

[13]Glennerster (2017) describes this back-and-forth process between researcher and partner about what to test.

[14]See also Glennerster (2017, p.189).

and planning how to integrate it into operation, and any disruptions to normal operation with attendant time costs for lower-level management and labor can be substantial and are often not fully covered by the experimenter.[15]

In sum, in the case of risky experiments, some of the same calculations are likely to be relevant for a site selecting into treatment autonomously and a site selecting into treatment via an experiment. We explore this argument in the next Section with a simplified Roy model examining selection into a risky treatment in both settings.

Implications of similar Bias–1. Building on the previous arguments, assume for the remainder of this Section that the Bias–1 of each setting exists and operates fairly similarly. In this case, both approaches yield biased estimates of the population average treatment effect, and suffer from a kind of selection bias that is often considered sufficient to relegate the observational approach to "non-causal" status.

The potential for Bias–1 in an observational study is an issue of internal validity and thus forefront in economists' minds. But the same kind of bias in experimental work is in the realm of external validity, and hence receives far less scrutiny. Further, the experimental approach achieves internal validity in part by shifting a key selection problem from the realm of internal to external validity; Bias–1 is not eliminated by the experiment, simply made external to the study. Thus, the overall superiority of the experimental approach seems partly a mirage. More broadly, the presumption that judgment ought to be based on internal validity properties alone appears unsupportable if key potential biases are shifted between the realms of internal and external validity in the methods under comparison.

Given the nature of the selection problem, the standard strategy to bolster external validity – replication of the same experiment in new sites – need not help. Every experiment that gets implemented would involve a site selecting into the experimental treatment and thus suffer from a similar bias.[16]

[15]See Heckman (1992, 2020) and Hotz (1992) for documentation of this concern in the context of a job training program in the US.

[16]See also Banerjee and Duflo (2009).

With respect to the broader question of "what works" in general, experimental evidence would not be obviously superior to observational. This is true despite the experimental approach's ability to eliminate Bias–2; two biases may not be worse than one, as we illustrate next.

4 Selection model

Assume that

$$Y_{ij} = \mu_j + U_{ij}, \quad j \in \{0,1\} \ ,$$

with $E(U_{ij}) = 0$, $j \in \{0,1\}$. The ATE parameter of interest is $\overline{\tau} = E(Y_{i1} - Y_{i0}) = \mu_1 - \mu_0$.

Departing from the standard Roy model, assume a site-specific component of the individual-i disturbance:

$$U_{ij} = \nu_{S_i j} + u_{ij}, \quad j \in \{0,1\} \ .$$

That is, ν_{sj} is a common component of U_{ij} for all individuals i in site s.

The individual component u_{ij} is not observed by the site-s DM, while the site-specific component is decomposed as follows:

$$\nu_{sj} = \nu_{sj}^O + \nu_{sj}^U, \quad j \in \{0,1\} \ ,$$

where ν_{sj}^O (ν_{sj}^U) is observable (unobservable) to the site-s DM. Let $\{\nu_{sj}^O\} \equiv (\nu_{s0}^O, \nu_{s1}^O)$ and $\{\nu_{sj}^U\} \equiv (\nu_{s0}^U, \nu_{s1}^U)$. Assume that for all $s \in \mathcal{S}$ and $j \in \{0,1\}$, u_{ij} is mean-zero and independent of $\{\nu_{sj}^O\}$ and $\{\nu_{sj}^U\}$, and $E(\nu_{sj}^U | \{\nu_{sj}^O\}) = 0$. Given that $E(U_{ij}) = 0$, these imply that $E(\nu_{sj}^O) = 0$, $j \in \{0,1\}$, $\forall s \in \mathcal{S}$.

Define $\overline{\tau}_s^O \equiv \nu_{s1}^O - \nu_{s0}^O$ and $\overline{\tau}_s^U \equiv \nu_{s1}^U - \nu_{s0}^U$. Then the (true) site-specific average treatment effect for site s is

$$\overline{\tau}_s \equiv E(\tau_i | S_i = s, \{\nu_{sj}^O\}, \{\nu_{sj}^U\}) = \mu_1 - \mu_0 + \nu_{s1}^O - \nu_{s0}^O + \nu_{s1}^U - \nu_{s0}^U \ = \overline{\tau} + \overline{\tau}_s^O + \overline{\tau}_s^U \ .$$

That is, each site's treatment effect is the population treatment effect $\overline{\tau}$ plus an observed and unobserved site-specific common effect.

The site-s DM makes decisions to maximize the expected outcomes of site-s individuals, knowing the model,[17] given what is observable, and thus with an expected site-s average treatment effect of

$$\overline{\tau}_s^{DM} \equiv E(\tau_i | S_i = s, \{\nu_{sj}^O\}) = \mu_1 - \mu_0 + \nu_{s1}^O - \nu_{s0}^O = \overline{\tau} + \overline{\tau}_s^O \ .$$

We examine two distinct scenarios. In one, all sites decide autonomously whether to take up the treatment. In the other, sites take up the treatment only when approached by a researcher offering to implement the treatment in an RCT framework. One can think of these two scenarios as one in which knowledge of the treatment product or policy diffused to everyone and the other in which it is only revealed by an experimenter. While mixed scenarios may be more realistic, these provide two simple benchmarks.

4.1 Autonomous Selection

In the first scenario, all sites know about and decide whether to undertake the treatment. As a baseline case, we assume that sites do not learn anything about the treatment's efficacy after undertaking it. This simplifies the analysis, and can be justified if sites do not know how to causally identify the treatment effect without an experimenter. An alternative interpretation is that the policy gets entrenched once adopted at the site.[18] Regardless, in this case treatment is a once-for-all decision with no learning, so site s adopts the policy iff

[17] This includes $\overline{\tau}$; one can easily extend the model for pessimistic or optimistic assessments of the average treatment effect. It also includes the distribution of $\{\nu_{sj}^U\}$.

[18] We conjecture that a similar but possibly weaker comparison result holds if instead sites learn their average treatment effects perfectly, i.e. if they learn ν_{sj}^U. The selection bias will be stronger in that case, as sites will be selected based on more accurate information. The degree of difference between the two cases would depend on the relative importance of ν_{sj}^O and ν_{sj}^U. However, assuming that sites can learn about the treatment autonomously would also complicate the decision to partner with an external experimenter.

$E(Y_{i1}|S_i = s, \{\nu_{sj}^O\}) > E(Y_{i0}|S_i = s, \{\nu_{sj}^O\})$, which is equivalent to

$$E\left(\tau_i|S_i = s, \{\nu_{sj}^O\}\right) > 0 \qquad \Longleftrightarrow \qquad \overline{\tau}_s^{DM} > 0 \qquad \Longleftrightarrow \qquad \overline{\tau} + \overline{\tau}_s^O > 0 \ . \tag{4}$$

Thus, $T_i = 1$ if $S_i \in \mathcal{S}_1 \equiv \{s \in \mathcal{S} : \overline{\tau}_s^O > -\overline{\tau}\}$ while $T_i = 0$ if $S_i \in \mathcal{S}_0 \equiv \mathcal{S} \setminus \mathcal{S}_1$.

Now consider collecting observational data on treatment status and outcomes of a random sample of individuals from the population and estimating the treatment effect using simple treated-untreated outcome comparisons. The bias in this approach is detailed in Section 3.1, equation 1. Given the simple selection pattern resulting from this model, we can solve for the two biases:

$$\begin{aligned} Bias\text{-}1 &\equiv E(\tau_i|S_i \in \mathcal{S}_1) - \overline{\tau} \\ &= E\left(\mu_1 - \mu_0 + \nu_{s1}^O - \nu_{s0}^O + \nu_{s1}^U - \nu_{s0}^U + u_{i1} - u_{i0} \mid S_i = s, \overline{\tau}_s^O > -\overline{\tau}\right) - \overline{\tau} \\ &= E\left(\nu_{s1}^O - \nu_{s0}^O \mid \overline{\tau}_s^O > -\overline{\tau}\right) \ , \qquad \text{and} \end{aligned} \tag{5}$$

$$\begin{aligned} Bias\text{-}2 &\equiv E[Y_{i0}|S_i \in \mathcal{S}_1] - E[Y_{i0}|S_i \in \mathcal{S}_0] \\ &= E\left(\mu_0 + \nu_{s0}^O + \nu_{s0}^U + u_{i0} \mid S_i = s, \overline{\tau}_s^O > -\overline{\tau}\right) \\ &\qquad - E\left(\mu_0 + \nu_{s0}^O + \nu_{s0}^U + u_{i0} \mid S_i = s, \overline{\tau}_s^O \leq -\overline{\tau}\right) \\ &= E\left(\nu_{s0}^O \mid \overline{\tau}_s^O > -\overline{\tau}\right) - E\left(\nu_{s0}^O \mid \overline{\tau}_s^O \leq -\overline{\tau}\right) \ , \end{aligned} \tag{6}$$

where the final equalities in both cases use the fact that the $\{\nu_{sj}^U\}$ and the $\{u_{ij}\}$ are mean-zero conditional on the $\{\nu_{sj}^O\}$. The total bias of the estimated treatment is

$$Bias\text{-}1 + Bias\text{-}2 = E\left(\nu_{s1}^O \mid \overline{\tau}_s^O > -\overline{\tau}\right) - E\left(\nu_{s0}^O \mid \overline{\tau}_s^O \leq -\overline{\tau}\right) \ .$$

4.2 Experimental Selection

In the second case, researchers must find a partner willing to engage in the experiment, i.e. to accept the treatment for a subset of individuals at its site. Site DMs can undertake

the treatment only if partnering with an experimenter, and they take into account both the short-run cost or benefit of the experiment in directly altering clients' payoffs and the long-run benefit of learning and thus potentially improving clients' payoffs in the future.

To allow for these dynamics, the period corresponding to the experimental treatment is distinguished from the post-experimental period. Discounting parameter $\delta \in (0,1)$ captures the relative weight on each period. The experiment is successful in producing useful information with probability $\phi \in (0,1)$; when successful, it reveals $\{\nu_{sj}^U\}$ and thus $\overline{\tau}_s$. After the experiment is run, the treatment is adopted permanently iff it is then expected to give higher average payoffs. A final parameter, $\lambda \in (0,1)$, captures exogenously the fraction of site s that is treated in the experiment.

Site-s DM accepts the experiment with its temporary treatment iff it is expected to give higher average payoffs, which is equivalent to the following inequality:[19]

$$\begin{aligned} E[Y_{i0} \,|\, \{\nu_{sj}^O\}\,] \quad &< \quad (1-\delta)\, E[\,\lambda Y_{i1} + (1-\lambda) Y_{i0} \,|\, \{\nu_{sj}^O\}\,] + \\ &\delta\phi\, E\left(max\left\{ E\left[Y_{i1} \,|\, \{\nu_{sj}^O\}, \{\nu_{sj}^U\}\right] , E\left[Y_{i0} \,|\, \{\nu_{sj}^O\}, \{\nu_{sj}^U\}\right] \right\} \,\middle|\, \{\nu_{sj}^O\}\right) + \\ &\delta(1-\phi)\, max\left\{ E\left[Y_{i1} \,|\, \{\nu_{sj}^O\}\right] , E\left[Y_{i0} \,|\, \{\nu_{sj}^O\}\right] \right\} . \end{aligned} \tag{7}$$

The left-hand side is the per-client present discounted value of refusing the experiment. The right-hand side is the analogous payoff for accepting the experiment, with the first term giving expected payoffs during the experiment, the second term featuring the forecast of expected post-experiment payoffs if the experiment is successful (revealing $\{\nu_{sj}^U\}$), and the third term featuring the forecast of post-experiment payoffs if the experiment fails to reveal anything. After subtracting the left-hand side term and simplifying the expectations, this inequality can be rewritten

$$(1-\delta)\lambda\,(\overline{\tau}+\overline{\tau}_s^O)+\delta(1-\phi)\, max\left\{\overline{\tau}+\overline{\tau}_s^O, 0\right\}+\delta\phi\, E\left(max\left\{\overline{\tau}+\overline{\tau}_s^O+\overline{\tau}_s^U, 0\right\} \,\middle|\, \{\nu_{sj}^O\}\right) > 0\,. \tag{8}$$

[19] For brevity we suppress the dependence of all expectations on $S_i = s$ here and in Inequality 8.

The first term is the site-s DM's expected benefit or cost to individual outcomes, relative to the status quo, of running the experiment. The second term captures the expected future gains relative to the status quo of making the optimal site-wide treatment choice when nothing has been learned from the experiment. The core of both of these terms, $\overline{\tau} + \overline{\tau}_s^O$, is identical to the one in Condition 4 for the autonomous case. The third term is similar to the second, except that the optimal site-wide treatment choice incorporates learning $\{\nu_{sj}^U\}$ from the experiment. This term captures the potential for a dynamic benefit from running the experiment; its core is weakly positive and greater than the second term's, and strictly so if learning $\{\nu_{sj}^U\}$ could alter the DM's decision.

In order to characterize and compare selection, we assume (A1) that all sites' $\{\nu_{sj}^O\}$ are drawn from a discrete joint distribution which does not allow the observable site-specific treatment effect to mirror the ATE exactly. That is,

$$\exists\, \epsilon > 0 \;\; s.t. \;\; \overline{\tau} + \overline{\tau}_s^O \notin (-\epsilon, \epsilon) \;, \;\; \forall s \in \mathcal{S} \;. \tag{A1}$$

We also assume boundedness of site unobservable effects $\{\nu_{sj}^U\}$:

$$\exists\, B^U > 0 \;\; s.t. \;\; |\nu_{s0}^U|, |\nu_{s1}^U| \leq B^U \;, \;\; \forall s \in \mathcal{S} \;. \tag{A2}$$

These assumptions lead to the following result:

Lemma 1. *Under assumptions A1 and A2, if*

a) Given treatment intensity λ, perceived informativeness of the experiment ϕ, and bound on site-level uncertainty B^U, impatience is sufficiently high (δ low enough), or

b) Given treatment intensity λ, level of patience δ, and bound on site-level uncertainty B^U, perceived informativeness of the experiment is sufficiently low (ϕ low enough), or

c) Given treatment intensity λ, level of patience δ, and perceived informativeness of the experiment ϕ, site-level uncertainty is sufficiently low (B^U low enough),

then assent to the experiment occurs iff $s \in \mathcal{S}_1 \equiv \{s \in \mathcal{S} : \overline{\tau}_s^O > -\overline{\tau}\}$.[20]

That is, under assumptions A1 and A2 and conditions a), b), or c) – referred to below as "Lemma 1 conditions" – selection into experimental treatment is identical to the autonomous selection into treatment of the previous section.

Consider site manager patience (condition a). If managers care little about the future, then all that matters is how the treatment affects site-s individuals today, i.e. the expected treatment effect, just as in the autonomous case. If concern for the future is stronger, selection into the experiment revolves around a possible tradeoff between short-run effects of running the experiment and long-run learning benefits.

Patience reflects a number of factors here. One is the degree of short-termism of the firm or NGO DM; depending on the nature of accountability to shareholders or donors, the ability to take short-term risk for long-term benefit may be small. Another factor is the duration of the experiment. If short-run results are informative and the experimental period is thus short, δ is larger; but if long-run results are preferred, the experimental period becomes longer and δ smaller.[21]

Consider next the perceived informativeness of the experiment (condition b). If it is thought to be relatively uninformative, then the decision to participate in the experiment reduces to a forecast of whether the treatment is worthwhile – as in the autonomous case. With some expected informativeness of the experiment, selection into the experiment revolves again around a possible tradeoff between short-run effects of running the experiment and

[20] *Proof.* Assent to the experiment occurs iff Inequality 8 holds. Given that its second and third terms are weakly positive, assent clearly occurs if $\overline{\tau} + \overline{\tau}_s^O > 0$, i.e. for all $s \in \mathcal{S}_1$. It remains to show that under conditions a), b), or c), assent is withheld at all other sites, i.e. those for which $\overline{\tau} + \overline{\tau}_s^O < -\epsilon$, for some $\epsilon > 0$ (using A1). For such sites, the negation of Inequality 8 simplifies to

$$\overline{\tau} + \overline{\tau}_s^O \leq -\frac{\delta\phi\, E\left(max\left\{\overline{\tau} + \overline{\tau}_s^O + \overline{\tau}_s^U, 0\right\} \mid \{\nu_{sj}^O\}\right)}{\lambda(1-\delta)} \equiv -X \ ,$$

say. The expectation term is bounded by 0 and B^U (using A2). Clearly if δ or ϕ is small enough (e.g. $\phi \leq \epsilon\lambda(1-\delta)/(\delta B^U)$), then $X \leq \epsilon$; while if $B^U < \epsilon$, then $X = 0$. It follows that under any of the three conditions, $\overline{\tau} + \overline{\tau}_s^O < -\epsilon$ guarantees that $\overline{\tau} + \overline{\tau}_s^O < -X$. ■

[21] Rosenzweig and Udry (2020) demonstrate that key parameters of interest can vary significantly over time, demonstrating the possibility that a short-run estimate may provide suboptimal information.

the potential for long-run learning benefits.

Perceived experimental informativeness depends in part on trust in the experimenter. The site DM may justifiably wonder to what degree the experimental team is organized, skilled, and unbiased. Also related is the durability of knowledge gained. The model assumes a fixed environment; in reality, environments and effects of policies may change over time. If the value of the knowledge gained through the experiment depreciates more rapidly, experimental informativeness drops.[22] Another issue from beyond the model is free-riding in learning. If the site DM believes that the experiment will be carried out by other sites if this site opts out, and that results readily extrapolate across sites, then the marginal information gain of this site's experiment will be believed to be low and current payoffs of the treatment will loom larger in the decision.[23]

Finally, the degree of uncertainty about site-specific potential outcomes also matters (condition c). If there is little site-specific treatment effect variation that is unobserved to the site-DM, i.e. ν_{s0}^U and ν_{s1}^U are negligible in magnitude, then there is little to learn from even a very successful experiment, and participation depends only on an initial forecast of whether the treatment is worthwhile. On the other hand, if there is a lot of uncertainty and the potential for high upside outcomes, the dynamic benefit from experimenting and learning the truth is high.[24]

Outside the Lemma, consider the size of the experimental treatment, λ. If only a small fraction of the site needs to be treated to gain significant knowledge, λ near 0, then the first term in inequality 8, capturing the short-run effect, gets negligible weight. If the amount of patience, perceived experimental informativeness, and residual site-level uncertainty are all non-negligible, the experiment will be assented to by every site, as the third term capturing

[22]This could be modeled via a probability of a new draw of $\{\nu_{sj}^U\}$ materializing in the post-experimental period; a higher such probability would directly map into lower experimental informativeness ϕ in the model as written.

[23]As modeled, there is no benefit to extrapolation across sites since each site knows the (population-level) ATE. Extrapolation would become useful if this assumption were changed to incorporate possibly incorrect subjective beliefs of site DMs about the ATE.

[24]In a model with risk aversion instead of risk neutrality, higher uncertainty could also raise expected short-run costs, leaving ambiguous overall implications.

learning benefits will dominate. Conversely, if a significant fraction of the site needs to be treated, due to the site itself being moderate in size or a preference against intra-site extrapolation, the current payoff term becomes a factor and selection may mirror the autonomous case.[25]

In summary, under Lemma 1 conditions, selection into treatment occurs identically via an experiment as through autonomous choice: $T_i = 1$ iff $S_i \in \mathcal{S}_1$. The bias can then be written, using equation 3 of Section 3.1 and equation 5:

$$Bias\text{-}1 \equiv E(\tau_i | S_i \in \mathcal{S}_1) - \overline{\tau} = E\left(\nu_{s1}^O - \nu_{s0}^O \mid \overline{\tau}_s^O > -\overline{\tau}\right) \ .$$

Here the total bias of the experimentally estimated treatment is upward, because sites opting into the experiment are positively selected based on their somewhat accurate forecast of treatment effect.

4.3 Comparing the Two Methodologies

The previous sections provide conditions under which selection into a risky experimental treatment is similar to autonomous selection into the treatment. The similarity in selection occurs, when it does, because the same forecasts of cost and benefit of the risky new policy or program feature prominently in decision-making in both contexts.

Even when selection into treatment is identical, the biases may differ significantly across the two approaches in this model. Recall that the total bias in the autonomous case is

$$E\left(\nu_{s1}^O \mid \overline{\tau}_s^O > -\overline{\tau}\right) - E\left(\nu_{s0}^O \mid \overline{\tau}_s^O \leq -\overline{\tau}\right) \tag{9}$$

[25]The model assumes DMs care only about the average experience of the site's individuals. But if a site had strong equity preferences, a decrease in λ could matter little; the DM could be unwilling to subject even a small minority of individuals to a risky experiment, despite the potential benefit to a large majority.

and in the experimental case under conditions a), b), or c) of Lemma 1 is

$$E\left(\nu_{s1}^{O} \mid \overline{\tau}_{s}^{O} > -\overline{\tau}\right) - E\left(\nu_{s0}^{O} \mid \overline{\tau}_{s}^{O} > -\overline{\tau}\right) \ . \tag{10}$$

The first term in each bias is identical, and results from *treated* outcomes being estimated from sites preferring to opt into the treatment, whether experimentally or autonomously. The second terms differ, because in the autonomous case the *untreated* outcomes are estimated from sites preferring to opt out of the treatment, while in the experimental case they are estimated from sites preferring to opt into treatment.

To illustrate further, we consider two simple cases.

Case 1: $\nu_{s1}^{O} = 0, \ \forall \ s$, while $\nu_{s0}^{O} \in \{\nu^{l}, \nu^{h}\}$, with $\nu^{l} < 0 < \nu^{h}$ and $|\overline{\tau}| < |\nu^{l}|, |\nu^{h}|$.

This case corresponds to heterogeneity in site-DM observed untreated outcomes, but none in treated outcomes. For example, the treatment could be an accounting and information system that is expected to achieve a certain level of efficiency for any firm (i.e. site); firms differ in the expected efficiency of their status quo accounting practices. Or, the treatment is expected to lead to a certain rate of loan default for any microfinance institution (i.e. site); MFIs differ in their status quo expected default rates.

Comparing the magnitudes of the bias in each case leads to the following Proposition:

Proposition 1. *Assume Condition a), b), or c) of Lemma 1 holds. The observationally estimated ATE is less biased if* $|\nu^{h}| < |\nu^{l}|$*, while the experimentally estimated ATE is less biased if* $|\nu^{l}| < |\nu^{h}|$*.*

Proof. Setting $\nu_{s1}^{O} = 0$ in equation 9 for the bias in the observational case gives that bias as $-E\left(\nu_{s0}^{O} \mid -\nu_{s0}^{O} \leq -\overline{\tau}\right)$, which equals $-\nu^{h}$. Setting $\nu_{s1}^{O} = 0$ in equation 10 for the bias in the experimental case – which applies under conditions a), b), or c) of Lemma 1 – gives that bias as $-E\left(\nu_{s0}^{O} \mid \nu_{s0}^{O} < \overline{\tau}\right)$, which equals $-\nu^{l}$. ■

This result holds because if ν^{l} is closer to zero (the mean of ν_{s0}^{O}), the experimental approach employs the more representative untreated counterfactual, sites preferring to opt

into treatment ($s \in \mathcal{S} : \nu_{s0}^{O} = \nu^{l}$); while if ν^{h} is closer to zero, the observational approach employs the more representative untreated counterfactual, sites preferring to opt out of treatment ($s \in \mathcal{S} : \nu_{s0}^{O} = \nu^{h}$).

Thus, an observational approach can produce better evidence on the ATE than an experimental one. There is no necessary advantage to the experimental approach. This would call into question, in a context like this, a benchmarking of observational results to experimental results (with respect to the population ATE); here both are biased, in different directions, and either can be more biased.

Why might the experimental approach, which at least eliminates one bias, Bias–2, still be inferior to the cross-sectional approach that eliminates neither bias? Adding more bias to a biased estimate can help or hurt. In this context, Bias–1 and Bias–2 counteract, so the addition of Bias–2 can reduce total bias.[26]

Note that this result does not rely on the observational study having greater scope than the experimental one.[27] The same result would clearly hold if the observational study included random samples of individuals from only a random sample of sites, and the RCT were replicated in as many (willing) sites as were represented in the observational study.

Proposition 1 holds under certain conditions. Under other conditions, e.g. if the experiment is somewhat informative and $\lambda \to 0$, selection into the experiment is universal and the experimental approach is unbiased for the ATE while the observational approach is biased. Thus, the point is not that observational methods are usually better than experimental, even in the case of risky experiments. It is instead that there is no obvious ranking; the nature of selection and treatment heterogeneity matter, and substantiating a claim to superiority of experimental methods in this context requires a statement about the parameters governing selection and heterogeneity.

In this setting, there is a summary statistic that sheds light on the relative strength of

[26]Further, the decomposition into Bias–1 and Bias–2 is somewhat arbitrary.

[27]Greater scope and representativeness of data is one reason observational studies have been argued to often have greater external validity (e.g. Dehejia, 2015).

each approach: the probability a random site would select into the experiment, call it α (see also Czibor et al., 2019). If $\alpha = 1$, the experimental approach is unbiased. But if only sites with $\nu_{s0}^O = \nu^l$ select into the experiment, as in the focal analysis above, it follows (from $E(\nu_{s0}^O) = 0$) that $\alpha = \nu^h/(\nu^h - \nu^l) \in (0, 1)$. The experimental bias magnitude can be written $|\nu^l| = (1 - \alpha)(\nu^h - \nu^l)$ while the observational bias magnitude can be written $|\nu^h| = \alpha(\nu^h - \nu^l)$.[28] Thus, the experimental bias is higher the fewer sites are willing to experiment (i.e. the lower α). If few enough sites are willing to experiment, i.e. α less than 1/2, then the experimental bias exceeds the observational bias. Applying this logic more broadly suggests that risky experiments where reluctance to participate is widespread and finding a willing partner is hard may be no more informative about the ATE than an analogous observational study, due to the significant possibility of selection bias.

Case 2: $\nu_{s0}^O = 0, \ \forall \ s$.

Here site-DM observed heterogeneity in untreated outcomes is ruled out, allowing only such heterogeneity in treated outcomes. For example, all firms (sites) are perceived to have similar accounting and information systems ex ante, but firm DMs differ in how well they anticipate the new treatment to fit with their firm's organization and skills. Or, MFIs have similar baseline loan default rates, but the treatment is expected to affect default rates differently at different MFIs.

Comparing the magnitudes of the bias in each case leads to the following Proposition:

Proposition 2. *Assume Condition a), b), or c) of Lemma 1 holds. The observationally estimated ATE and the experimentally estimated ATE have equal bias.*

Proof. Setting $\nu_{s0}^O = 0$ in equation 1 for the bias in the observational case and in equation 3 for the bias in the experimental case – which applies under conditions a), b), or c) of Lemma 1 – gives both biases as $E\left(\nu_{s1}^O \mid \nu_{s1}^O > -\overline{\tau}\right)$. ■

Any bias in the observational case comes from treated sites having larger expected treat-

[28] These follow from the facts that $\alpha = \nu^h/(\nu^h - \nu^l)$, $1 - \alpha = -\nu^l/(\nu^h - \nu^l)$, and (from the Proof of Proposition 1) the experimental bias is $-\nu^l$ while the observational bias is $-\nu^h$.

ment effects – this is why they opted in. But the same bias exists in the experimental case – the same sites with larger expected treatment effects are the ones willing to participate in the experiment.

We have also considered mixtures of Cases 1 and 2, where ν_{s0}^O and ν_{s1}^O each have two possible values and can be correlated across sites. Unsurprisingly, the comparison becomes more complicated, but either approach can be less biased depending on the specifics of the parameters.

5 Extensions

5.1 Biased expectations

In the baseline model, site DMs have unbiased beliefs about the site-specific and population ATE. Here we assume instead that site DMs may have biased beliefs.

We modify the baseline model by assuming that site DMs make decisions based on subjective expectations and, letting $\mathcal{E}$ denote subjective expectations, that $\mathcal{E}(\overline{\tau}_s^U|\{\nu_{sj}^O\}) = -\pi$ for all $s \in \mathcal{S}$. The remainder of the model is unchanged, including that $E(\nu_{sj}^U|\{\nu_{sj}^O\}) = E(\overline{\tau}_s^U|\{\nu_{sj}^O\}) = 0$ and that all other subjective expectations correspond to reality. Note that π measures pessimism; the site-s DM expected site-specific average treatment effect is

$$\overline{\tau}_s^{DM} \equiv \mathcal{E}(\tau_i|S_i = s, \{\nu_{sj}^O\}) = \overline{\tau} + \overline{\tau}_s^O - \pi$$

and DMs believe the population average treatment effect to be $\overline{\tau} - \pi$. (DMs are optimistic if $\pi < 0$.)

Now autonomous selection into treatment occurs iff $\overline{\tau}_s^{DM} > 0$, i.e. $T_i = 1$ iff $S_i \in \mathcal{S}_1' \equiv \{s \in \mathcal{S} : \overline{\tau}_s^O > -\overline{\tau} + \pi\}$. Selection into experimental treatment occurs iff

$$(1-\delta)\lambda\left(\overline{\tau}+\overline{\tau}_s^O-\pi\right)+\delta(1-\phi)\,max\left\{\overline{\tau}+\overline{\tau}_s^O-\pi,0\right\}+\delta\phi\,\mathcal{E}\left(max\left\{\overline{\tau}+\overline{\tau}_s^O+\overline{\tau}_s^U,0\right\}\,\middle|\,\{\nu_{sj}^O\}\right)>0\,,$$

modifying Inequality 8. A nearly identical version of Lemma 1 continues to hold and provide conditions under which selection into treatment occurs at exactly the same sites ($\mathcal{S}_1'$) in the experimental case as in the autonomous case. Under those conditions, we can write the total bias in the observational case as

$$E\left(\nu_{s1}^{O} \mid \overline{\tau}_s^{O} > -\overline{\tau} + \pi\right) - E\left(\nu_{s0}^{O} \mid \overline{\tau}_s^{O} \leq -\overline{\tau} + \pi\right)$$

and in the experimental case as

$$E\left(\nu_{s1}^{O} \mid \overline{\tau}_s^{O} > -\overline{\tau} + \pi\right) - E\left(\nu_{s0}^{O} \mid \overline{\tau}_s^{O} > -\overline{\tau} + \pi\right) \ ,$$

modifying Equations 9 and 10. The first term in each bias is the same; both reflect that sites opting into treatment – the sites on which the treated counterfactual is based – are a more selected sample the more groundlessly pessimistic site DMs are.[29] The second terms, representing the untreated counterfactual, differ. In the observational case, the relevant sites are the sites that opt out, and these are *less* selected (more typical) the more pessimism there is; while in the experimental case, the relevant sites for the untreated counterfactual are the sites opting into treatment, and these are *more* selected under greater pessimism.

Thus, greater pessimism among site DMs about treatment will tend to tilt the bias comparison in favor of the observational case, since its untreated counterfactual comes from opt-outs, which are more representative under more pervasive pessimism. This suggests a limited ability of RCT methodology to persuade a skeptical set of decision makers. Conversely, greater optimism about treatment will likely tilt the bias comparison in favor of the experimental case, since opt-ins are then increasingly representative.

[29] That is, higher π makes the condition governing the conditional expectation more stringent.

5.2 Autonomous learning

In the autonomous case of the baseline model, site DMs do not learn about the treatment's effects over time, but make a once-and-for-all adoption decision. Here we assume instead that site DMs do learn about the treatment if they adopt it for some period of time, and that they may discard it after learning about it.

If sites can learn autonomously, then the decision has the same structure as in the experimental case of the baseline model: expected current payoffs of adopting the treatment for some subset of the site, and expected long-run payoffs after learning and then choosing optimally. In fact, we assume the same learning structure as in the experimental case, captured by Inequality 7, except with potentially different parameters capturing the fraction treated in the learning phase, λ^A, and the probability of successfully learning, ϕ^A. Simplification leads to a condition nearly identical to Inequality 8 governing whether a site selects treatment:

$$(1-\delta)\lambda^A\left(\overline{\tau}+\overline{\tau}_s^O\right)+\delta(1-\phi^A)\,max\left\{\overline{\tau}+\overline{\tau}_s^O,0\right\}+\delta\phi^A\,E\left(max\left\{\overline{\tau}+\overline{\tau}_s^O+\overline{\tau}_s^U,0\right\}\;\middle|\;\{\nu_{sj}^O\}\right)>0\;.$$

There are now two straightforward ways for site selection to be the same in experimental and autonomous cases: it is identical if $\lambda=\lambda^A$ and $\phi=\phi^A$, or if Lemma 1 conditions hold for both experimental and autonomous parameters.

There is a further stage of selection that may be relevant in the autonomous case with learning, however, as some sites that opt into treatment for the trial period may opt out of treatment after learning. Whether this selection affects observational estimation depends on whether population outcomes are measured during the early phase, while learning is happening, or during the mature phase when learning is complete.[30] We thus discuss four cases, based on the two ways for initial selection to be similar and whether the population is studied during the early phase or the mature phase.

I) Consider first early-phase study and identical selection due to Lemma 1 conditions.

[30]Of course, the experimental approach also allows sites to opt in or out after the trial period, but the data are typically collected while treatment is in place.

In this case, the selection is the same as that analyzed in Sections 4.1 and 4.2, and all the analysis and comparisons of Section 4.3 apply unchanged. Thus, allowing for autonomous learning does not necessarily affect the results, as selection may still be based on initial estimates of site-specific treatment effects in both cases.

II) Next, consider mature-phase study and identical selection due to Lemma 1 conditions. In this case, the researcher is observing $T_i = 1$ only for sites that initially opted in *and* that did not later opt out after learning. That is, $T_i = 1$ iff $S_i \in \mathcal{S}'_1$, where

$$\mathcal{S}'_1 \equiv \{ s \in \mathcal{S} : \overline{\tau}^O_s > -\overline{\tau} \ \ \& \ \ \overline{\tau}^O_s + \overline{\tau}^U_s > -\overline{\tau} \} \ .$$

Modifying equations 5 and 6, the total bias in the observational estimation can be written

$$Bias_{Obs} = Bias\text{-}1 + Bias\text{-}2 = E\left(\nu^O_{s1} + \nu^U_{s1} \mid s \in \mathcal{S}'_1\right) - E\left(\nu^O_{s0} + \nu^U_{s0} \mid s \in \mathcal{S}'_0\right) \ ,$$

where $\mathcal{S}'_0 \equiv \mathcal{S} \setminus \mathcal{S}'_1$. For comparison, a modified equation 8 gives the bias in the experimental estimation as[31]

$$Bias_{Exp} = E\left(\nu^O_{s1} + \nu^U_{s1} \mid \overline{\tau}^O_s > -\overline{\tau}\right) - E\left(\nu^O_{s0} + \nu^U_{s0} \mid \overline{\tau}^O_s > -\overline{\tau}\right) \ .$$

The first terms in each bias are identical in the baseline case, while here the first-term bias is more acute in the observational case since the ν^O_{s1} and ν^U_{s1} term are more positively selected. It is harder to know what happens to the second-term bias in the observational case – the ν^O_{s1} term is less positively selected, while the ν^U_{s1} term is more positively selected. The following examples illustrate the tradeoffs.

Case 1′: $\nu^O_{s1} = \nu^U_{s1} = 0, \ \forall \ s$; $\nu^O_{s0} \in \{\nu^l, \nu^h\}$, with $\nu^l < 0 < \nu^h$ and $|\overline{\tau}| < |\nu^l|, |\nu^h|$; and $\nu^U_{s0} \in \{\nu^L, 0, \nu^H\}$, with $\nu^L < -|\overline{\tau}| - \nu^h$ and $\nu^H > |\overline{\tau}| - \nu^l$. We also assume statistical independence between ν^O_{s0} and ν^U_{s0}.

[31] Here ν^U_{s1} and ν^U_{s0} are explicitly included to make the comparison clearer; they are omitted from equation 8 because they are mean-zero conditional on the $\{\nu^O_{sj}\}$.

Under these assumptions, which expand on Case 1 of Section 4.3,

$$\mathcal{S}_1' = \{s \in \mathcal{S} : \nu_{s0}^O = \nu^l \ \ \& \ \ \nu_{s0}^U \neq \nu^H \ \} \qquad \text{and} \qquad \mathcal{S}_0' = \{s \in \mathcal{S} : \nu_{s0}^O = \nu^h \ \ | \ \ \nu_{s0}^U = \nu^H \ \} \ ,$$

and the observational bias is

$$Bias_{Obs} = -[a\nu^h + (1-a)(\nu^H + \nu^l)] \ ,$$

where $a \in (0,1)$. Specifically, letting $p^h \equiv Prob(\nu_{s0}^O = \nu^h)$ and $p^H \equiv Prob(\nu_{s0}^U = \nu^H)$, $a = p^h/[p^h + (1-p^h)p^H]$. The experimental bias remains $Bias_{Exp} = -\nu^l$.

Two results follow. First, the observational bias in this case with learning may be smaller or larger than in the baseline case with no autonomous learning. This is clear since the bias in that case is ν^h, while here it is a convex combination of ν^h and $\nu^H + \nu^l$, either of which may be larger (assumptions guarantee that both terms exceed $|\overline{\tau}|$, but nothing more). Thus, autonomous learning can push the bias comparison in either direction. Second, even if the observational bias is larger here than in the baseline case, it may still exceed the experimental bias, though under stronger conditions than those of Proposition 1.

Case 2′: $\nu_{s0}^O = \nu_{s0}^U = 0, \ \forall \ s$; $\nu_{s1}^O \in \{\nu^l, \nu^h\}$, with $\nu^l < 0 < \nu^h$ and $|\overline{\tau}| < |\nu^l|, |\nu^h|$; and $\nu_{s1}^U \in \{\nu^L, 0, \nu^H\}$, with $\nu^L < -|\overline{\tau}| - \nu^h$ and $\nu^H > |\overline{\tau}| - \nu^l$. We also assume statistical independence between ν_{s1}^O and ν_{s1}^U.

Under these assumptions, which expand on Case 2 of Section 4.3,

$$\mathcal{S}_1' = \{s \in \mathcal{S} : \nu_{s1}^O = \nu^h \ \ \& \ \ \nu_{s1}^U \neq \nu^L \ \} \qquad \text{and} \qquad \mathcal{S}_0' = \{s \in \mathcal{S} : \nu_{s1}^O = \nu^l \ \ | \ \ \nu_{s1}^U = \nu^L \ \} \ ,$$

and the observational bias is

$$Bias_{Obs} = \nu^h + \frac{p^H}{1-p^L}\nu^H \ ,$$

$p^H \equiv Prob(\nu^U_{s1} = \nu^H)$ and $p^L \equiv Prob(\nu^U_{s1} = \nu^L)$. This is greater than the experimental bias, which remains at $Bias_{Exp} = \nu^h$. However, the difference can be small if the outlier values ν^H and ν^L are relatively rare. Further, it is clear that in combinations of Cases $1'$ and $2'$ may see either approach result in less bias.

III) Consider next early-phase study and identical selection due to similar parameters ($\phi = \phi^A$, $\lambda = \lambda^A$) rather than Lemma 1 conditions. In this case, selection is identical in autonomous and experimental settings, but different for both than in the baseline case where the set of sites opting in was $\mathcal{S}_1 = \{s \in \mathcal{S} : \overline{\tau}^O_s > -\overline{\tau}\}$. Here one can write the set of sites opting in as

$$\mathcal{S}'_1 = \{s \in \mathcal{S} : \overline{\tau}^O_s > -\overline{\tau} \ \mid \ \text{Condition 8 holds } \} \ ,$$

which is a superset of $\mathcal{S}_1$. Now one can compare the two biases as follows:

$$Bias_{Obs} = E\left(\nu^O_{s1} + \nu^U_{s1} \mid s \in \mathcal{S}'_1\right) - E\left(\nu^O_{s0} + \nu^U_{s0} \mid s \in \mathcal{S} \setminus \mathcal{S}'_1\right) \ ,$$

and

$$Bias_{Exp} = E\left(\nu^O_{s1} + \nu^U_{s1} \mid s \in \mathcal{S}'_1\right) - E\left(\nu^O_{s0} + \nu^U_{s0} \mid s \in \mathcal{S}'_1\right) \ .$$

As usual, the first terms in each bias are identical, while the second terms are different due to measuring the untreated counterfactual with the opt-outs (observational) or opt-ins (experimental). The difference with the baseline case is that the opt-in set here is bigger ($\mathcal{S}'_1$), so the selection is likely to be less severe in the experimental case than in the observational case. In the extreme case where nearly all sites opt in because of the learning benefits, the second-term vanishes in the experimental case and gets relatively large in the observational case. Away from extreme cases, autonomous learning should tilt the bias comparison against the observational case, though it can still be the less biased.

IV) The case of mature-phase study and identical selection due to similar parameters is the most complicated. It combines the forces at play in cases II) and III). Without delving into greater detail, we conjecture that the overall outcome is likely to tilt the comparison

against the observational approach while leaving open the possibility that it remains less biased than the experimental approach.

On the whole, then autonomous learning seems to favor the experimental case, but may not be decisive in the comparison.

6 Discussion and Conclusion

"Risky" experiments – experiments that involve partnering with a site to adopt a risky treatment – may suffer from a canonical selection bias with respect to the population ATE. Perhaps because the experimental approach shifts it from the realm of internal to external validity, this potential selection bias appears to be largely ignored in the experimental literature. Arguably, it should not be; the sensitivity to selection bias that is regularly applied to observational studies seems potentially just as applicable to risky RCT studies.

We argue that experimental estimates in this context should be explicitly framed as (ideally) unbiased for the ATE of the site where the experiment was carried out, or as unbiased for the ATE of willing sites. Ideally, the site selection process should be documented, including identity of sites approached and negotations with partner sites that determined the specific policy tested.

Given that identification is of the ATE of willing sites, it should be recognized that skeptical sites (NGOs, firms, states, etc.), in particular those that would not have opted into the experimental treatment, are on standard econometric ground in treating evidence from even much-replicated RCTs as "non-causal" – at least as a guide for their own decisions. Related, the potential for RCTs to substantially add to general knowledge about high-stakes treatments may be relatively limited.

While the paper focuses on a stark participation margin, selection bias may also enter more subtly through negotated participation. For example, if an NGO consents to an experiment but has refused certain treatments and altered others during negotiation, concern

about selection into treatment seems no less relevant. As a minimum, ideally researchers report on all negotiations with all NGOs so the reader can gauge the nature of selection into the experiment;[32] the challenge is to standardize this reporting to keep it from being subjective or incomplete.

This paper does not provide evidence on where and to what degree selection biases are operative in RCTs. Conversely, there appears to be no evidence elsewhere that these biases are *not* operative in risky experiments. This is an avenue for further research, especially since the bias may differ significantly across contexts (e.g. with the level of the treatment's risk). While evidence is lacking, as with observational studies the conservative approach is to presume that this kind of selection may be operative, or to be clear about assumptions ruling it out. To this end, the modeling in this paper of factors that can accentuate the selection bias – lack of patience, prevalence of knowledge free-riding, and so on – can provide a basis for identification assumptions in the RCT context.

Are there solutions to this potential selection issue in risky experiments? One possibility is to compensate the partner site for participation, to the extent that virtually any site would have selected into the experiment. To bring this out of the realm of guesswork, a process could be documented in which a standardized experimental offer is given to a large number of potential partner sites, and assent to participation is nearly universal among all offered the experiment. In some contexts, such a liquidity infusion could affect results. An alternative to lump-sum compensation is to insure the partner site against downside risk of participation; however, this raises a concern about distortion of site incentives, i.e. moral hazard.

Another solution is to adapt the techniques used to deal with individual-level selection into treatment – estimating ITT or LATE parameters – to address site-level selection. Estimating these parameters at the site level would be possible after extending experimental treatment offers to a number of potential partner sites, possibly randomized, and tracking

[32]See Belot and James (2016) for further commentary on selection-relevant reporting.

outcomes at participating and non-participating sites – a potentially costly approach. Further, a LATE parameter would represent sites selecting into treatment due to being offered the experiment, and thus would potentially involve selection issues similar to those modeled in this paper.[33]

A third approach would be to interpret estimates as bounds by signing the bias of the ATE. Roy model biases are typically positive, as they are in this paper's model. If so, an estimate of zero may be informative that there truly is no positive effect, given the upward bias in the estimate.

A final approach is simply to frame RCT results as unbiased for the ATE for *willing* sites, and discuss what willingness may imply about selection in the given setting. The observational approach seems helpful as an analogy in this discussion: to what degree would selection bias be of concern in an analogous observational study, and might the same kind of selection bias be operative in the experimental setting?[34]

Ultimately, the standard experimental approach does not give obviously more unbiased estimates for a broader population than an observational study, in a site-based, risky-treatment context. This is true even with diverse replication. Of course, the experimental approach may do better, but the assumptions required to be confident in its smaller bias seem substantive and worth making explicit.

[33]Specifically, never-taking sites would be on standard ground treating the evidence as "non-causal". Gechter and Meager (2022) demonstrate the significant possibilities of this kind of approach, though with a somewhat different definition of site selection, based on researcher's choice of technique rather than site manager's decision to adopt or pilot a novel treatment.

[34]Specifically, for a given experiment, imagine that treatment was instead autonomously selected across a number of sites, and an cross-site observational study was carried out. Would such a study be vulnerable to the (internal validity) critique of site selection bias? If so, on what grounds are we confident a similar site selection bias is not rendering results from an RCT similarly biased (via external validity) for the ATE?

References

[1] Hunt Allcott. Site selection bias in program evaluation. *Quarterly Journal of Economics*, 130(3):1117–1165, August 2015.

[2] Isaiah Andrews and Emily Oster. A simple approximation for evaluating external validity bias. *Economics Letters*, 178:58–62, May 2019.

[3] Susan Athey and Guido W. Imbens. The econometrics of randomized field experiments. In *Handbook of Economic Field Experiments*, volume 1, pages 73–140. North-Holland, 2017.

[4] Abhijit V. Banerjee, Sylvain Chassang, Sergio Montero, and Erik Snowberg. A theory of experimenters: Robustness, randomization, and balance. *American Economic Review*, 110(4):1206–1230, April 2020.

[5] Abhijit V. Banerjee, Sylvain Chassang, and Erik Snowberg. Decision theoretic approaches to experimental design and external validity. In *Handbook of Economic Field Experiments*, volume 1, pages 141–174. North-Holland, 2017.

[6] Abhijit V. Banerjee and Esther Duflo. The experimental approach to development economics. *Annual Review of Economics*, 1(1):151–178, 2009.

[7] Michèle Belot and Jonathan James. A new perspective on the issue of selection bias in randomized controlled field experiments. *Economics Letters*, 124(3):326–328, 2014.

[8] Michèle Belot and Jonathan James. Partner selection into policy relevant field experiments. *Journal of Economic Behavior & Organization*, 123:31–56, March 2016.

[9] Eszter Czibor, David Jimenez-Gomez, and John A List. The dozen things experimental economists should do (more of). *Southern Economic Journal*, 86(2):371–432, 2019.

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

[11] Rajeev Dehejia. Experimental and non-experimental methods in development economics: A porous dialectic. *Journal of Globalization and Development*, 6(1):47–69, 2015.

[12] Greg Fischer and Dean Karlan. The catch-22 of external validity in the context of constraints to firm growth. *American Economic Review, Papers & Proceedings*, 105(5):295–299, 2015.

[13] Michael Gechter. Generalizing the results from social experiments: Theory and evidence from Mexico and India. Working Paper, July 18, 2022.

[14] Michael Gechter and Rachael Meager. Combining experimental and observational studies in meta-analysis: A debiasing approach. Working Paper, June 20, 2022.

[15] Rachel Glennerster. The practicalities of running randomized evaluations: Partnerships, measurement, ethics, and transparency. In *Handbook of Economic Field Experiments*, volume 1, pages 175–244. North-Holland, 2017.

[16] James J. Heckman. Randomization and social policy evaluation. In Irwin Garfinkel and Charles F. Manski, editors, *Evaluating Welfare and Training Programs*, pages 201–230. Harvard University Press, 1992.

[17] James J. Heckman. Randomization and social policy evaluation revisited. In Florent Bédécarrats, Isabelle Guérin, and François Roubaud, editors, *Randomized Control Trials in the Field of Development: A Critical Perspective*, pages 304–330. Oxford University Press, 2020.

[18] James J. Heckman, Robert J. LaLonde, and Jeffrey A. Smith. The economics and econometrics of active labor market programs. In *Handbook of Labor Economics*, volume 3, pages 1865–2097. Elsevier, 1999.

[19] James J. Heckman and Edward J. Vytlacil. Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In *Handbook of Econometrics*, volume 6, pages 4875–5143. Elsevier, 2007.

[20] V. Joseph Hotz. Designing an Evaluation of the Job Training Partnership Act. In Irwin Garfinkel and Charles F. Manski, editors, *Evaluating Welfare and Training Programs*, pages 76–114. Harvard University Press, 1992.

[21] V. Joseph Hotz, Guido W. Imbens, and Julie H. Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1-2):241–270, March-April 2005.

[22] Guido W. Imbens. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2):399–423, June 2010.

[23] Anup Malani. Patient enrollment in medical trials: Selection bias in a randomized experiment. *Journal of Econometrics*, 144(2):341–351, 2008.

[24] Charles F. Manski. *Public policy in an uncertain world: Analysis and decisions*. Harvard University Press, 2013.

[25] Rachael Meager. Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, January 2019.

[26] Jonathan Morduch. The disruptive power of RCTs. In Florent Bédécarrats, Isabelle Guérin, and François Roubaud, editors, *Randomized Control Trials in the Field of Development: A Critical Perspective*, pages 108–125. Oxford University Press, 2020.

[27] Seán M. Muller. Causal interaction and external validity: Obstacles to the policy relevance of randomized evaluations. *World Bank Economic Review*, 29:S217–S225, 2015.

[28] Lant Pritchett and Justin Sandefur. Context matters for size: Why external validity claims and development practice do not mix. *Journal of Globalization and Development*, 4(2):161–197, 2014.

[29] Lant Pritchett and Justin Sandefur. Learning from experiments when context matters. *American Economic Review, Papers & Proceedings*, 105(5):471–475, 2015.

[30] Martin Ravallion. Should the Randomistas (continue to) rule? In Florent Bédécarrats, Isabelle Guérin, and François Roubaud, editors, *Randomized Control Trials in the Field of Development: A Critical Perspective*, pages 47–78. Oxford University Press, 2020.

[31] Mark R. Rosenzweig and Christopher Udry. External validity in a stochastic world: Evidence from low-income countries. *The Review of Economic Studies*, 87(1):343–381, 2020.

[32] Eva Vivalt. How much can we generalize from impact evaluations? *Journal of the European Economic Association*, 18(6):3045–3089, December 2020.