How and when is a "risky" RCT more biased than a naive observational approach?

Christian Ahlin*

August 2024

Abstract

We compare the potential biases in estimating a population average treatment effect (ATE) using an experimental approach and an observational approach that naively compares treated and untreated outcomes. 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. 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 and provide conditions in a Roy model under which the ATE is estimated with greater bias using the experimental approach rather than the observational. For example, if the ATE is low, many sites are reluctant to participate, or site managers are pessimistic about the ATE, even a much-replicated RCT may produce more biased evidence than 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 an open question in the area of microfinance, such as whether group lending 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 for empirical testing may be simply to compare outcomes of microfinance institutions ("MFIs") that use the different approaches, for example group- and non-group based lenders. Many would consider this option significantly inferior to an experimental approach: 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 widely thought to generate the single most credible form of empirical evidence on causal impact, likely due to their ability to produce an unbiased estimate of the average treatment effect ("ATE") under minimal assumptions and because they allow the researcher control over the mechanism for assignment to treatment (e.g., Imbens, 2010, Athey and Imbens, 2017).

This paper argues that in certain natural settings, the researcher's control over the assignment mechanism can be hampered much as it would be in a decentralized setting, leading the experimental approach to produce a more biased estimate of the population ATE than a simple observational approach. The settings we focus on involve the population partitioned into subsets called "sites" – MFIs, states, NGOs, firms, banks, etc. Sites have two key features: within-site correlation, so that average treatment effects may differ across sites; and a single decision-maker responsible for determining treatment. Each site's decision-maker is benevolent and decides whether to undertake treatment, having partial knowledge of how well the treatment works in context. The decision-maker's objective coincides with the outcome of interest to the researcher. Crucially, the treatment is new to the site, and *risky*: it has a real chance of appreciably worsening outcomes at the site, as in 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, bias may plague 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" status.

The RCT approach involves a researcher partnering with a particular site to test the treatment experimentally. Randomization would ideally eliminate expected 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 would not threaten internal validity of the study, but would be classified as an external validity issue. Still, if non-zero it would bias the experimental estimate of the population ATE, even after many replications.

This paper argues that both approaches may face a fundamentally similar challenge: "Bias–1", which in both the observational and experimental cases is the selection bias stemming from the willingness of sites to select into the treatment, autonomously or experimentally. It is known that Bias–1 is zero in the experimental approach if willingness to assent to experimental treatment is uncorrelated with site treatment effects (Hotz et al., 2005, Allcott, 2015). But it would seem natural in many contexts to allow for optimizing site-level decision-makers taking their own forecasts of the treatment effect into account in assenting to treatment, whether autonomously or experimentally, and especially when the treatment carries significant risk. If so, both the observational approach and the experimental approach may suffer from a similar, canonical selection bias with respect to the population ATE. The experimental approach does not eliminate this potential bias, but rather shifts it from the realm of internal validity to external validity where it receives less scrutiny.

Still, the nature of site selection may be more different than similar in autonomous and experimental cases, in which case "Bias–1" may not be analogous across the two. Further, even if selection is similar, the experimental approach may be considered preferable because it at least gets rid of one potential bias, if not both: "Bias–2", via randomization.

To explore these issues further, we use a standard Roy model of selection into treatment in both autonomous and experimental settings. Our main departure from the standard setup is in modeling site-level selection and allowing for site-specific effects, partially observed by site decision-makers. In the microcredit example, this corresponds to MFI managers having an informative prior about whether group lending, equity contracts, or frequent meetings will work in their context. We find that under some conditions – e.g. short-termism of the site manager, limited trust in the experimenting team or the experimental process, knowledge free-riding problems, or significant cost in partnering with the experimenter – the selection pattern and thus "Bias–1" are indeed similar across the autonomous and the experimental settings.

Having characterized selection in both cases, the model enables comparison of the total bias in each case, Bias–1 in the experimental case and Bias–1 + Bias–2 in the observational case. We find that the observational approach can easily produce a less biased estimate of the ATE than the experimental approach. In particular, under similar selection and plausible assumptions on the joint distribution of site-specific effects, if the ATE is negative, or positive and not too large, the observational approach is less biased, while if the ATE is sufficiently large and positive, the experimental approach is less biased. Since willingness to engage in the experiment increases with the ATE, a corollary is that if few sites are willing to participate, the experimental approach is more biased, while if the probability of finding a willing site is high enough ($> 3/4$ in our i.i.d. standard normal simulation), the experimental approach is less biased. These results suggest that "hard-to-sell" experiments may not be worth selling.

Why does the experimental approach not dominate by eliminating at least one bias,

"Bias–2"? The reason is that Bias–1 and Bias–2 may counteract, so that the sum of both biases may be closer to zero than Bias–1 by itself. Thus we have a setting in which bias from external validity issues in an experimental approach – due to site selection into experimentation – can be more severe than the biases due to lack of internal validity in a naive observational approach. This casts doubt on the approach of categorically valuing internal validity over external validity in the quest for general knowledge.

In extensions, we show that even as site decision-makers' information about their site-specific ATE diminishes, the bias comparison can be unaffected. Diminished information about the site ATE makes the experiment closer to randomly assigned and thus closer to unbiased. But it also makes the experiment less necessary as naive observational comparisons are less biased, as autonomous treatment status also becomes more randomly assigned. We also show that the argument extends to site decision-makers with biased expectations, and that under reasonable conditions, pessimism (optimism) about the ATE tilts the bias comparison in favor of the observational (experimental) approach.

In sum, given the inability to escape selection issues that plague the observational approach in more obvious ways, a "risky" experiment – even a much-replicated one – may produce no better evidence on the population ATE than a naive observational approach. These results hold in a simple, standard model, demonstrating that nothing extraordinary is necessary for an experimental approach to underdeliver.

The experimental approach can unbiasedly estimate the site-specific ATE here, which in some cases may be the appropriate target parameter (Deaton and Cartwright, 2018). But the general interest appeal of the experimental approach arguably comes from a claim (sometimes implicit) to create insights into what works or how things work more broadly, across contexts (e.g., see Imbens, 2010, p. 417), and for this the site-specific ATE is inadequate.

The experimental approach can also produce an unbiased estimate of the ATE among sites that are willing to opt into the experiment (a Treatment on the Treated parameter). 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 sites unwilling to experiment may well abound, and a presumption that by virtue of randomization RCTs deliver more credible evidence for them than an analogous correlational study appears unfounded. Given the logic of site selection, site decision-makers that would not have agreed to the experiment are on standard econometric ground in doubting the implications of RCT evidence for themselves.[1] Policymakers who ignored this issue and advocated or imposed the treatment among sites that would not have been initially interested in the experiment would encounter in our model a predictable "voltage drop" in scaling up.

The paper also cautions against the idea that RCTs are most useful in testing out innovative modes of operation or techniques (e.g. Morduch, 2020) toward the accumulation of best practices. Unless the innovations are likely to be trivial or positive in impact – i.e. unless the experiment is non-"risky" – site selection possibilities give rise to 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 analysis does point to several possible remedies to the experimental site selection issue. First, if sites can be compensated so that the likelihood of selecting into the experiment is sufficiently high – without altering treatment effects – the selection issue disappears. Second, given even a discrete instrument, e.g. a randomly assigned site-level bonus for participation, one can test for external validity (e.g., Brinch et al., 2017): if site-level ATE estimates do not vary with the bonus payment, this is evidence against harmful site selection. Third, by taking a stand on the selection model, one might sign the bias of the experimental estimate and use a bounding argument. As is typical with Roy models, the experimental bias is positive in our case, so finding a non-positive estimate may be convincing evidence

[1]This perspective thus provides a caveat to the finding that the randomized approach is especially suited for convincing an adversarial audience (Banerjee et al., 2020). Their analysis applies to the site-specific ATE; ours differs in incorporating cross-site external validity. Banerjee et al. (2017) do deal with external validity and provide an 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 is complementary and differs in focusing on canonical selection issues.

against a positive ATE.

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 uses a Roy model for selection into treatment in both cases, and combines results with the statistical model to provide conditions under which each approach is more biased. Section 5 explores extensions to the selection model. Section 6 concludes, including with a discussion of solutions to the issues here and directions for further research. All proofs are in the Appendix.

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 pointed out 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, Athey and Imbens, 2017) and are likely to be practically solveable by replication across diverse sites. The current paper addresses these responses in showing that replication does nothing to remove a key bias, and that in some plausible cases, external validity issues lead to more damaging biases in experimental approaches than in observational.

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 and empirically demonstrates that bias with respect to the population parameter can arise due to systematic selection of sites into experimental participation, and makes 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 Deaton (2010, p.445) and Muller (2015). Similar arguments around site selection bias can be found in Heckman and Vytlacil (2007), Banerjee and Duflo (2009), Manski (2013), Pritchett and Sandefur (2013), Dehejia (2015), Fischer and Karlan (2015), Wydick (2016), 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".

We aim to explore, expand, and formalize the reasoning of these contributions regarding issues researchers must consider when generalizing from experiments, in a richer theoretical framework focused on site selection. Our focus is also narrower: not experiments in general, but the class of "risky", site-based experiments that involve adopting a new treatment. It is with these types of experiments that Roy-model selection – directly connected to effectiveness of the treatment – at the site level seems most difficult to assume away. Thus, while the literature tends to focus on incidental reasons for site selection into experimentation,[2] 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 calculations, 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. This paper's

[2]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).

embedding of both internal and external validity into a common model with a clearly specified parameter of interest allows for the juxtaposition and aggregation of biases – related to external and internal validity separately – across different types of studies. 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 lexicographically prioritizing internal validity over external validity in evaluating empirical methods can be misguided.[3]

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.[4]

The adoption of Roy model logic for selection into experiments 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. 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 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 feasible solutions to site-based selection.

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

[4]Thus we believe the paper follows up Deaton's (2010, p.445) point that "without an analysis of the two biases, it is impossible to conclude which estimate is better".

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. Invoking these kinds of assumptions undermines the case for RCTs being categorically superior for causal identification, as others have argued.

Several studies find ways to empirically assess site selection bias or other external validity issues. Allcott (2015) 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 in his data from an observational estimate (using within-site time variation rather than the across-site differences that 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,[5] we view it as helping to interpret empirical results from different methods and, in particular, highlighting the kinds of experiments that may in their typical deployment offer little improvement over basic observational methods in generating credible *general* knowledge.

Overall, the paper builds on a number of well-known ideas and further develops their

[5] Arguably, none of the studies cited in this paragraph involves a clear-cut example of a "risky" experiment.

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 "sites", indexed by $s \in \{1, ..., M\}$, interpreted as states, locales, NGOs, firms, banks, MFIs, or contexts. The set of individuals at site s is $\mathcal{P}_s$, where $\{\mathcal{P}_1, ..., \mathcal{P}_M\}$ 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, individual 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 effect in site s, the "sATE": $\overline{\tau}_s \equiv E(\tau_i | S_i = s)$. The assumptions guarantee that $E(\overline{\tau}_s) = \overline{\tau}$.[6]

Second, treatment status of individuals within a site is determined by a site-level decision-maker, or "DM".[7] 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.[8]

[6]The expectation in the previous sentence is over $\mathcal{P}$ (individuals), while the expectation in this sentence is over $\mathcal{S}$ (sites). We use both throughout, aided by the assumption of equal site populations.

[7]Thus a school district with multiple schools or an MFI operating in multiple neighborhoods would each count as a single site, if the decision to undertake the treatment is made at the school-district or MFI-level.

[8]To focus on site-level selection, we abstract from individual-level selection into treatment and assume all individuals in a treated site are treated. This may be a reasonable assumption in the microfinance case, e.g., where an MFI offers take-it-or-leave-it policies to its clients and may change certain features of its product without altering its clientele significantly. Individual-level selection would create problems for both experimental and observational approaches in identifying the ATE; in the experimental approach, by possibly unfortunate necessity the focal parameter would likely become an ITT or LATE. A full analysis is beyond

3.1 Observational Approach

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 $S_i \in \mathcal{S}_0$.

Assume that treatment status T_i and its associated outcome $Y_i(T_i)$ are observed for a random sample of individuals i in $\mathcal{P}$. A simple observational approach compares sampled outcomes 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 into the policy tends to differ from the treatment effect in sites that opt out. This bias disappears if all sites have the same treatment effect, $\overline{\tau}_s = \overline{\tau}$, $\forall s$, or more generally if sites' treatment effects are uncorrelated with their adoption of the policy, $E(\tau_i|S_i \in \mathcal{S}_1) = E(\tau_i|S_i \in \mathcal{S}_0)$. Bias–2 arises if sites that opt in and sites that opt out would tend to have different outcomes without the policy. The possibility for either bias is widely seen as preventing simple correlational estimates from being interpreted as "causal".[9]

the scope of this paper, but similar issues to those raised here would be relevant.

[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).

3.2 Experimental Approach

Assume the researcher partners with the willing site-s DM to assign treatment to a 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 within 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 identified within the scope of the study to the population parameter of interest.

It is known[10] that Bias–1 is zero if there is no site-specific heterogeneity in treatment effects, and more generally, that it is zero in expectation if the probability of a site participating in the experiment is uncorrelated with its treatment effect. Thus, it is zero in expectation if site s is randomly chosen from all sites, since $E(\hat{\tau}_{rct,s}) = \overline{\tau}_s$ and $E(\overline{\tau}_s) = \overline{\tau}$. 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.

[10]See Allcott (2015), Muller (2015), as well as Heckman et al. (1999) and Heckman and Vytlacil (2007).

Less well understood is how reasonable this assumption is, and what the consequences are if it does not hold. It is in jeopardy because sites often have a say in whether they are experimented upon and how, and thus site participation may not be random.[11] 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, if sites are selected randomly subject to willingness to experiment, for the average treatment effect of sites willing to engage in the 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 draw broader policy implications.[12]

3.3 Comparing the Approaches

"Bias–1" of the observational case – $E(\tau_i | S_i \in \mathcal{S}_1) - \overline{\tau}$ (equation 1) – is not identical to "Bias–1" of the experimental case – $E(\tau_i | S_i \in \mathcal{S}_1^E) - \overline{\tau}$ (equation 3) – but conceptually related. Bias–1 in the observational case is related to sites' decision to opt into the policy autonomously; 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, opting in requires at least some individuals at the site to undergo treatment, and optimizing site DMs presumably decide in part based on how they expect the treatment to affect these individuals. That is, the framework above suggests that a standard, Roy-model selection

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

[12] See also Imbens (2010, p.417). Experimental studies seem to be typically framed, published, and cited for insights that generalize beyond the context studied; conversely, it seems uncommon for an RCT study to foreground inability to speak "causally" beyond the sATE or the ATE for *willing* sites.

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 when they are.

Is Bias–1 similar in the two cases? Bias–1 is presumptively treated as a concern in simple observational studies, since estimation relies on variation in the potentially endogenous decision of whether to adopt the policy.

Bias–1 may not be a concern in experiments where participation is not subject to consent (e.g., online retail firms experimentally varying prices or web layout, researchers submitting fabricated resumes) or where consent may be considered a foregone conclusion (e.g., unconditional cash transfers to households, budgetary support to a government agency). Site selection is then controlled by the experimenter, who may select the site(s) at random, eliminating Bias–1.

Site selection into the experiment may also not be a concern 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.[13] While willingness to have one's standard product evaluated can differ across potential experimentees, it may do so in a way that differs substantially from selection into a new treatment, the focus here.

This paper restricts attention to RCTs in which the site DM consents to adopt a new product or policy variation that involves non-negligible 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 not uncommon in development economics, for example, where researchers often attempt to convince an NGO or firm or government agency to test a 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

[13]Many RCTs fit into this category, e.g. the six microcredit impact evaluations discussed in Banerjee et al. (2015).

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, productivity, 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,[14] 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.[15] 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.

Some experiments are risky even if the possible measured treatment effects are confined within a positive range, since 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, 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

[14]Glennerster (2017) describes this back-and-forth process between researcher and partner about what to test; see also Banerjee and Duflo (2009).

[15]That experiment-implementing partners have useful knowledge of the local context is argued by Glennerster (2017, p.189).

experimenter.[16]

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 Roy model examining selection into a risky treatment in both settings.

Implications of similar Bias–1. Building on the previous arguments, assume for now that the Bias–1 of each setting exists and operates fairly similarly. In this case, both approaches yield potentially biased estimates of the population average treatment effect, and both 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 appears to receive less scrutiny. Further, it seems fair to say that 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, but simply made external to the study. If so, is the presumed superiority of the experimental approach partly a mirage? More broadly, the idea that judgment ought to be based lexicographically on internal validity properties appears unsupportable if key potential biases are shifted between the realms of internal and external validity in the methods being compared.

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 implemented would involve a site selecting into the experimental treatment and thus suffer from a similar bias.[17]

With respect to the broader question of "what works" in general, experimental evidence

[16]See Heckman (1992, 2020) and Hotz (1992) for documentation of this concern in the context of a job training program in the US, and Glennerster (2017).

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

would not be obviously superior to observational. This is true despite the experimental approach's ability to eliminate Bias–2, 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 unobservable:

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

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 decision maker (DM), while the site-specific component is decomposed as follows to allow for partial observability by the site DM:

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

where ν_{sj}^O (ν_{sj}^U) is observable (unobservable) to the site-s DM. No component of U_{ij} is observable to the researcher. 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 $j \in \{0,1\}$, u_{ij} is mean-zero and independent of $\{\nu_{S_ij}^O\}$ and $\{\nu_{S_ij}^U\}$, and ν_{sj}^U is mean-zero and independent of $\{\nu_{sj}^O\}$.[18] Thus $E(\nu_{sj}^O) = 0$, $j \in \{0,1\}$.

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

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

[18] It may be natural to think of ν_{sj}^O and ν_{sj}^U as positively correlated; however, one can interpret ν_{sj}^O as incorporating the expectation of the unobserved component conditional on $\{\nu_{sj}^O\}$, and ν_{sj}^U as the conditionally mean-zero residual.

That is, each site's average treatment effect is the population average treatment effect $\overline{\tau}$ plus observed and unobserved site-specific common effects, τ_s^O and τ_s^U. We assume boundedness of the expected truncated unobserved part of the site treatment effect:

$$\exists\, B^U > 0 \quad s.t. \quad E\left(max\left\{\tau_s^U, 0\right\}\right) \leq B^U \ . \tag{A1}$$

The site-s DM makes decisions to maximize the expected outcomes of site-s individuals, knowing the model,[19] given what is observable to the DM, and thus with an expected site-s average treatment effect of

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

We examine two distinct scenarios. In one, all sites decide autonomously whether to take up the new treatment. In the other, sites take up the new 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 more realistically describe some settings, these provide two simple benchmarks.

4.1 Autonomous Selection

In the first scenario, all sites know about and decide whether to undertake the treatment autonomously. As a baseline case, we assume that sites do not learn anything about the treatment's efficacy after undertaking it. This simplifies the analysis,[20] 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.

In this case treatment is a once-for-all decision with no learning, so site s adopts the

[19] This includes includes the distribution of $\{\nu_{sj}^U\}$ and the ATE $\overline{\tau}$; Section 5.2 extends the model for incorrect beliefs about the ATE and $\{\nu_{sj}^U\}$.

[20] Section 5.3 relaxes this assumption.

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

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

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

Now assume $\mathcal{A}_0$ and $\mathcal{A}_1$ are non-empty and 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:

$$\begin{aligned} \textit{Bias}\text{-}1 &= E(\tau_i|S_i \in \mathcal{A}_1) - \overline{\tau} \\ &= E\left(\mu_1 - \mu_0 + \nu_{S_i1}^O - \nu_{S_i0}^O + \nu_{S_i1}^U - \nu_{S_i0}^U + u_{i1} - u_{i0} \mid \tau_{S_i}^O > -\overline{\tau}\right) - \overline{\tau} \\ &= E\left(\nu_{s1}^O - \nu_{s0}^O \mid \tau_s^O > -\overline{\tau}\right) \ , \qquad \text{and} \end{aligned} \tag{5}$$

$$\begin{aligned} \textit{Bias}\text{-}2 &= E(Y_{i0}|S_i \in \mathcal{A}_1) - E(Y_{i0}|S_i \in \mathcal{A}_0) \\ &= E\left(\mu_0 + \nu_{S_i0}^O + \nu_{S_i0}^U + u_{i0} \mid \tau_{S_i}^O > -\overline{\tau}\right) \\ &\qquad - E\left(\mu_0 + \nu_{S_i0}^O + \nu_{S_i0}^U + u_{i0} \mid \tau_{S_i}^O \leq -\overline{\tau}\right) \\ &= E\left(\nu_{s0}^O \mid \tau_s^O > -\overline{\tau}\right) - E\left(\nu_{s0}^O \mid \tau_s^O \leq -\overline{\tau}\right) \ , \end{aligned} \tag{6}$$

where the final equalities in both cases use the fact that $\{\nu_{sj}^U\}$ and $\{u_{i0}, u_{i1}\}$ are mean-zero conditional on $\{\nu_{sj}^O\}$. The total bias of the estimated treatment is thus

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

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; after the experiment they are free to permanently adopt or discard the treatment. 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. Discount factor $\delta \in (0,1)$ captures the fraction of weight placed on the post-experimental period. Parameter $\lambda \in (0,1)$, captures exogenously the fraction of site s that is treated in the experiment. The experiment is successful in producing useful information with probability $\phi \in (0,1)$, independently of all other variables; when successful, it reveals $\{\nu^U_{sj}\}$, and thus $\overline{\tau}_s$. A final parameter, $k \geq 0$, captures the site's per capita fixed cost of coordinating with the experimenter during the experimental period. This reflects the idea that the partnering organization bears some of the operational costs of experimentation – management costs, disruption costs, and others as discussed in Section 3.3 – even if the researcher covers costs of, e.g., training and treatment.[21]

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:[22]

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

The left-hand side is the per-client present discounted value of refusing the experiment. The

[21] A generalization would allow for costs of adopting the treatment in the autonomous case; insights from this model extend to that case, with k interpreted as the cost difference.

[22] For brevity we suppress the conditioning of all expectations in this inequality on $S_i = s$.

right-hand side is the analogous payoff for accepting the experiment, with the first term reflecting the cost of participating in the experiment, only paid during the experimental period; the second term giving expected payoffs during the experiment; the third term featuring the forecast of expected post-experiment payoffs if the experiment is successful (revealing $\{\nu_{sj}^U\}$); and the fourth term featuring the forecast of post-experiment payoffs if the experiment fails to reveal anything. The third and fourth terms reflect the anticipation that the treatment will be adopted iff it appears better after the experiment is run, successfully or not. After subtracting the left-hand side term and simplifying the expectations, this inequality can be rewritten

$$(1-\delta)\lambda\left(\overline{\tau}+\tau_s^O\right)+\delta(1-\phi)\,max\left\{\overline{\tau}+\tau_s^O,0\right\}+\delta\phi\,E\left(max\left\{\overline{\tau}+\tau_s^O+\tau_s^U,0\right\}\right)>(1-\delta)k\ , \quad (8)$$

where the expectation is over the unknown τ_s^U. The first term is the site-s DM's expected benefit or cost to individual potential 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, $\overline{\tau}+\tau_s^O$, 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.

We next show that under some conditions, selection into treatment is the same – asymptotically or exactly – as in the autonomous case. The conditions are of two types: first, even with no cost of participation ($k=0$), parameters can be such that the value or expectation of future gains from experimental learning are negligible; second, non-negligible expected gains from learning may be counterbalanced by the cost of participation.

Recall that $\mathcal{A}_1 \equiv \{s \in \mathcal{S} : \tau_s^O > -\overline{\tau}\}$. Further, define $\mathcal{A}_1^{+\epsilon} \equiv \{s \in \mathcal{S} : \tau_s^O > -\overline{\tau}-\epsilon\}$ and $\mathcal{A}_1^{-\epsilon} \equiv \{s \in \mathcal{S} : \tau_s^O > -\overline{\tau}+\epsilon\}$. Note that if $\epsilon > 0$, $\mathcal{A}_1^{-\epsilon} \subseteq \mathcal{A}_1 \subseteq \mathcal{A}_1^{+\epsilon}$.

Lemma 1. *A) Assume $k = 0$. Under assumption A1 and for any $\epsilon > 0$, if*

a) Given treatment intensity λ, perceived informativeness of the experiment ϕ, and bound on site-level uncertainty B^U,[23] *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 effectiveness 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 information is sufficiently good (B^U low enough),

then assent to the experiment occurs iff $s \in \mathcal{S}_1^E$, where $\mathcal{S}_1^E$ satisfies $\mathcal{A}_1 \subseteq \mathcal{S}_1^E \subseteq \mathcal{A}_1^{+\epsilon}$.

B) Define $\hat{K} \equiv \frac{\delta\phi}{1-\delta} E\left(max\left\{\tau_s^U, 0\right\}\right)$. If $k = \hat{K}$, then assent to the experiment occurs iff $s \in \mathcal{A}_1$. If $k > \hat{K}$ (resp., if $k < \hat{K}$), there exists an $\epsilon > 0$ such that assent to the experiment occurs iff $s \in \mathcal{A}_1^{-\epsilon}$ (resp., iff $s \in \mathcal{A}_1^{+\epsilon}$).

Part A of the Lemma shows that under assumption A1, selection into experimental treatment is asymptotically identical to the autonomous selection into treatment of the previous section if patience, experimental informativeness, or site-level information is sufficiently low; this would lead to $\mathcal{S}_1 = \mathcal{A}_1 \approx \mathcal{S}_1^E$.

Consider site manager patience (condition a). If managers care little about the future, then their decision is based on how the treatment affects site-s individuals in the short run, i.e. the expected treatment effect, just as in the autonomous case. 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.[24] 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.[25]

Consider next the perceived effectiveness of the experiment (condition b). If it is thought

[23] Recall that B^U is the upper bound on $E(max\{\tau_s^U, 0\})$ (see Assumption A1), a measure of site-level lack of information that can be remedied by the experiment.

[24] On the concern of short-termism, see Glennerster (2017).

[25] 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.

to have low chances of generating useable knowledge, then the decision to participate in the experiment reduces to an ex ante forecast of whether the treatment is worthwhile – as in the autonomous case. 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.[26] 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.[27]

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.[28]

The above conditions (a-c) ensure that the expected value from learning is sufficiently small compared to other concerns. But negligible expected learning value is not necessary for identical selection to occur, as part B of Lemma 1 establishes. The quantity $\hat{K}$ defined there can be interpreted as the value of learning at a site that is indifferent to the treatment

[26]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.

[27]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, as in the extension of Section 5.2.

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

without learning. The Lemma thus says that if organizational costs of participating in the experiment exactly balance these perceived learning benefits ($k = \hat{K}$), then selection into experimental treatment is identical to autonomous selection into treatment: $\mathcal{S}_1^E = \mathcal{A}_1 (= \mathcal{S}_1)$.

If participation costs are higher than learning benefits ($k > \hat{K}$), selection into the experiment is more acute than autonomous selection into treatment: sites must strictly prefer the treatment without learning, since learning benefits fall short of participation costs at the relevant point of indifference. The reverse holds if costs are lower ($k < \hat{K}$). This points to another factor that would make selection a more acute issue in experiments: high organizational costs borne by the experimental partner.

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, and if participation is costless ($k = 0$), the experiment will be assented to by every site, as the third term capturing learning benefits will dominate. Conversely, if a significant fraction of the site needs to be treated, due to the site itself being modest in size or a preference against intra-site extrapolation, the current payoff term becomes a factor and selection may mirror the autonomous case even when $k = 0$.[29] Costs of participation ($k > 0$) strengthen the argument; in particular, $k = \hat{K}$ guarantees that selection is identical to the autonomous case regardless of λ.

In summary, Lemma 1 provides conditions under which selection into treatment occurs identically, or virtually identically, via an experiment as through autonomous choice. Assuming then that in the experimental case $T_i = 1$ iff $S_i \in \mathcal{A}_1$ and that $\mathcal{A}_1$ is non-empty, the

[29]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. An important factor here would be the ability to compensate experimental subjects.

bias can then be written, using equation 3 of Section 3.1 and equation 5:

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

This experimental bias 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 an 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 new policy or program feature prominently in decision-making in both contexts.

Even when selection into treatment is identical, the biases of the two approaches differ. This is true in the statistical model of Section 3 without imposing the Roy selection model of Section 4. In the observational case, using Equation 1,

$$\begin{aligned} Bias_{Obs} &= \underbrace{E(\tau_i|S_i \in \mathcal{S}_1) - \overline{\tau}}_{Bias-1} + \underbrace{E\left(Y_{i0}|S_i \in \mathcal{S}_1\right) - E\left(Y_{i0}|S_i \in \mathcal{S}_0\right)}_{Bias-2} \\ &= E\left(Y_{i1}|S_i \in \mathcal{S}_1\right) - E\left(Y_{i0}|S_i \in \mathcal{S}_0\right) - \overline{\tau} \ , \end{aligned}$$

while in the experimental case, using Equation 3 and $\mathcal{S}_1 = \mathcal{S}_1^E$,

$$Bias_{Exp} = \underbrace{E(\tau_i|S_i \in \mathcal{S}_1^E) - \overline{\tau}}_{Bias-1} + \underbrace{0}_{Bias-2} = E\left(Y_{i1}|S_i \in \mathcal{S}_1\right) - E\left(Y_{i0}|S_i \in \mathcal{S}_1\right) - \overline{\tau} \ .$$

These biases are identical except for the term capturing untreated values (Y_{i0}), since untreated values are estimated from sites opting out of treatment ($S_i \in \mathcal{S}_0$) in the observational approach but from sites opting into treatment ($S_i \in \mathcal{S}_1$) in the experimental approach. (Treated values (Y_{i1}) are estimated from sites opting into treatment in both approaches.)

Imposing the selection model of this Section and identical selection across contexts (see Lemma 1), the total bias in the autonomous case with non-empty $\mathcal{S}_0$ and $\mathcal{S}_1$ becomes (see end of Section 4.1)

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

and in the experimental case with non-empty $\mathcal{S}_1^E = \mathcal{S}_1$ (see end of Section 4.2)

$$Bias_{Exp} = E\left(\nu_{s1}^O \mid \tau_s^O > -\overline{\tau}\right) - E\left(\nu_{s0}^O \mid \tau_s^O > -\overline{\tau}\right) \ . \tag{10}$$

Again, the first term in each bias is identical since *treated* outcomes are estimated from sites preferring to opt into the treatment under both methodologies. The second terms differ because the *untreated* outcomes are estimated from sites opting out of (into) the treatment in the autonomous (experimental) case.

To facilitate comparison of the above biases, assume that ν_{s0}^O and ν_{s1}^O are distributed with no mass points over support $\mathbb{R}$. For $x \in \mathbb{R}$, define

$$K_1(x) \equiv E(\nu_{s1}^O|\tau_s^O > -x), \qquad K_0(x) \equiv E(\nu_{s0}^O|\tau_s^O \leq -x), \qquad \tilde{K}_0(x) \equiv E(\nu_{s0}^O|\tau_s^O > -x) \ , \tag{11}$$

so that $Bias_{Obs} = K_1(\overline{\tau}) - K_0(\overline{\tau})$ and $Bias_{Exp} = K_1(\overline{\tau}) - \tilde{K}_0(\overline{\tau})$ (see equations 9 and 10). Note that

$$\lim_{x\to\infty} K_1(x) = \lim_{x\to-\infty} K_0(x) = \lim_{x\to\infty} \tilde{K}_0(x) = 0 \ ,$$

since ν_{s1}^O and ν_{s0}^O are mean-zero. Since $K_1(x) = E(\nu_{s1}^O|\nu_{s1}^O > \nu_{s0}^O - x)$, it is reasonable to assume that $K_1'(x) < 0$; extending this logic supports the assumption that

$$K_1'(x) < 0, \qquad K_0'(x) > 0, \qquad \tilde{K}_0'(x) > 0 \ . \tag{A2}$$

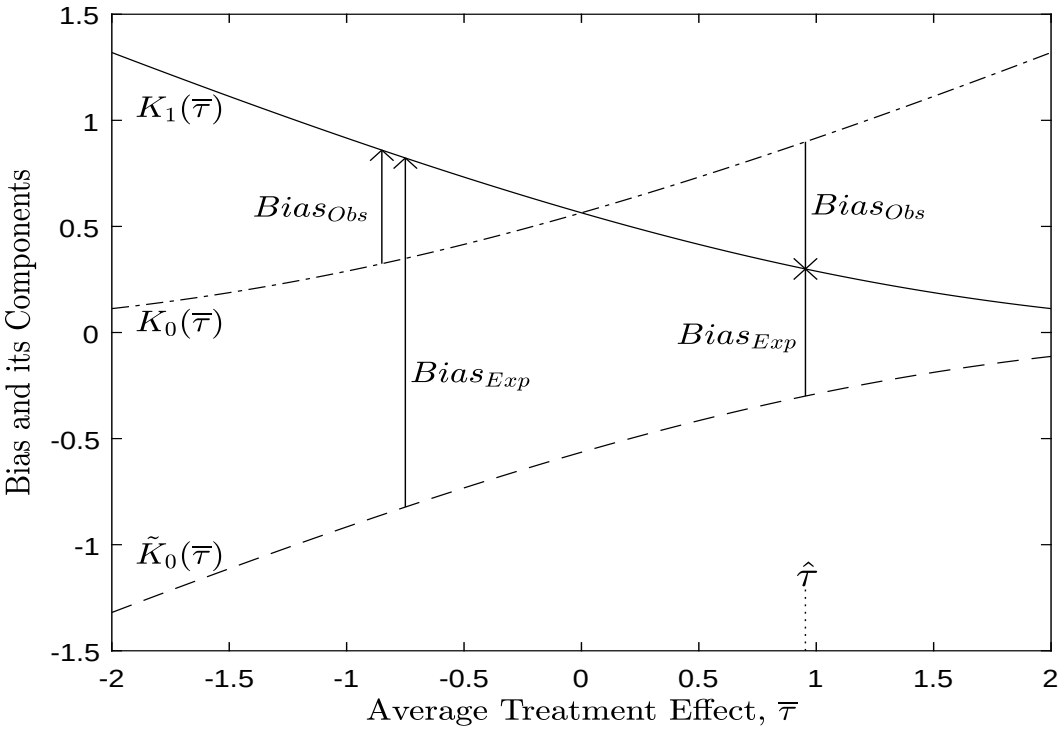


Figure 1: Displays bias of the observational and experimental approaches for different values of the average treatment effect, $\overline{\tau}$. Note that since $\mathcal{S}_1 = \mathcal{S}_1^E = \mathcal{A}_1$, $Bias_{Obs} = K_1(\overline{\tau}) - K_0(\overline{\tau})$ and $Bias_{Exp} = K_1(\overline{\tau}) - \tilde{K}_0(\overline{\tau})$. Here ν_{s0}^O and ν_{s1}^O are iid standard normal, and the experimental approach is less biased iff $\overline{\tau} > \hat{\tau} = \sqrt{2}\,\Phi^{-1}(3/4) \approx 0.95$.

Under this assumption,[30] we have the following result.

Proposition 1. *Under Assumption A2, and assuming identical selection in the observational and experimental approaches, i.e.* $\mathcal{S}_1 = \mathcal{S}_1^E = \mathcal{A}_1 \equiv \{s \in \mathcal{S} : \tau_s^O > -\overline{\tau}\}$*, there exists a* $\hat{\tau}$ *such that the observational bias is smaller in magnitude if* $\overline{\tau} < \hat{\tau}$ *and the experimental bias is smaller in magnitude if* $\overline{\tau} > \hat{\tau}$*. Assuming further that* ν_{s0}^O *and* ν_{s1}^O *are iid, then* $\hat{\tau} > 0$*.*

That is, either approach can result in greater bias when selection is the same (conditions guaranteeing which are provided in Lemma 1). If the site-specific errors ν_{s0}^O and ν_{s1}^O are iid, the observational approach is less biased for $\overline{\tau}$ negative, or positive and not too large.

Figure 1 illustrates the Proposition's logic. It graphs $K_1(\overline{\tau})$, $K_0(\overline{\tau})$, and $\tilde{K}_0(\overline{\tau})$ for $\{\nu_{sj}^O\}$

[30] One can show that, if ν_{s1}^O and ν_{s0}^O are iid, these functions are *coarsely* increasing or decreasing in the directions assumed, since $\lim_{x\to-\infty} K_1(x) = \lim_{x\to\infty} K_0(x) = \infty$ and $\lim_{x\to-\infty} \tilde{K}_0(x) = -\infty$; but local non-monotonicity is possible. It is also generally possible for at least some of the derivatives to be reversed, e.g. if ν_{s1}^O and ν_{s0}^O are strongly positively correlated.

iid standard normal. The observational and experimental biases can be read from the differences between the relevant functions, as the true $\overline{\tau}$ varies. Clearly both biases are decreasing in $\overline{\tau}$; the experimental bias declines toward zero, while the observational bias declines through zero and is always smaller (i.e. closer to $-\infty$, since $\tilde{K}_0(\overline{\tau}) < K_0(\overline{\tau})$). Thus at some cutoff value for $\overline{\tau}$, here $\hat{\tau} \approx 0.95$, the observational bias is as negative as the experimental bias is positive; to the left (right), the observational (experimental) bias is smaller in magnitude. Since ν_{s0}^O and ν_{s1}^O are iid, the observational bias is zero at $\overline{\tau} = 0$, guaranteeing that the cutoff $\hat{\tau}$ is positive.

The intuition is that the observational approach estimates untreated outcomes from sites opting out of treatment (leading to bias $K_0(\overline{\tau})$) while the experimental approach estimates untreated outcomes from sites opting into treatment (leading to bias $\tilde{K}_0(\overline{\tau})$). Sites opting out are more representative of the population when $\overline{\tau}$ is low, since most sites opt out; when $\overline{\tau}$ is high, sites opting in are more representative.[31]

Thus, a naive observational approach can produce better evidence on the ATE than an experimental one. 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.[32]

Note that this result does not rely on the observational study having greater scope than the experimental one. 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, $k = 0$, and $\lambda \to 0$, then selection into the experiment is universal and the experimental approach is unbiased for the ATE. Thus, the point is not that

[31]The reasoning is slightly more complex: what matters is not which approach estimates untreated outcomes with less bias, but which estimates them with a bias closest to the bias in estimated treated outcomes, $K_1(\overline{\tau})$. Since $K_1(\overline{\tau}) > 0$, this tilts the comparison toward the observational approach.

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

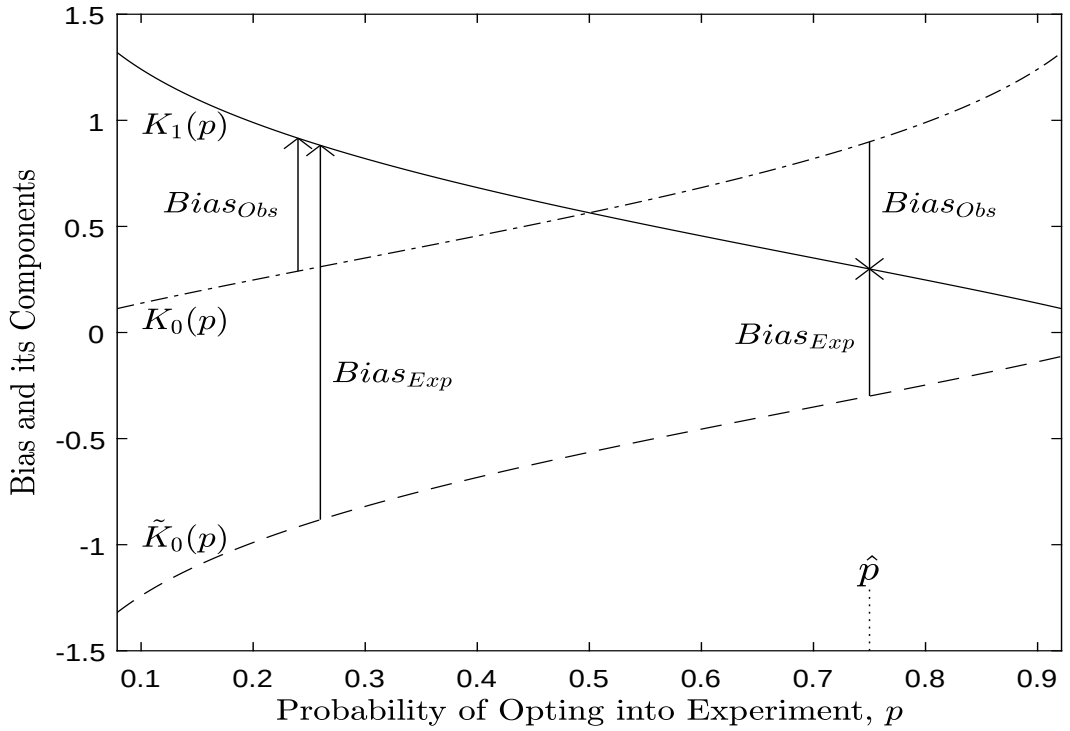


Figure 2: Displays bias of the observational and experimental approaches for different probabilities that a site selects into (autonomous or experimental) treatment, p, as $\overline{\tau}$ varies. Note that since $\mathcal{S}_1 = \mathcal{S}_1^E = \mathcal{A}_1$, $Bias_{Obs} = K_1(p) - K_0(p)$ and $Bias_{Exp} = K_1(p) - \tilde{K}_0(p)$. Here ν_{s0}^O and ν_{s1}^O are iid standard normal. The experimental approach is less biased iff $p(\overline{\tau}) > \hat{p} = 3/4$.

observational methods are usually better than experimental in the case of risky experiments, but that there is no obvious ranking when selection is endogenized; the nature of selection and treatment heterogeneity matters.

In this setting, there is a summary statistic that sheds light on the relative strength of each approach: the probability a random site would select into the experiment (see also Czibor et al., 2019). Specifically, let $p(\overline{\tau})$ be this probability, i.e. that a randomly selected site s is in $\mathcal{A}_1$. Equivalently, $p(\overline{\tau}) = Prob(\tau_s^O > -\overline{\tau})$.

Corollary 1. *Under Assumption A2, and assuming identical selection in the observational and experimental approaches, i.e.* $\mathcal{S}_1 = \mathcal{S}_1^E = \mathcal{A}_1 \equiv \{s \in \mathcal{S} : \tau_s^O > -\overline{\tau}\}$*, there exists a* $\hat{p}$ *such that the observational bias is smaller if* $\overline{\tau}$ *is such that* $p(\overline{\tau}) < \hat{p}$ *and the experimental bias is smaller if* $\overline{\tau}$ *is such that* $p(\overline{\tau}) > \hat{p}$*.*

Figure 2 replicates Figure 1, but rescaling the horizontal axis to be the probability a site selects into the experiment, as $\overline{\tau}$ varies. Note that if few enough sites are willing to experiment, here if $p < \hat{p} = 3/4$, 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 significant selection bias.

5 Extensions

5.1 Limited site-DM knowledge

In the baseline model, site DMs have information about the site-specific component of the treatment effect; they observe $\{\nu_{sj}^O\}$. This drives site selection into treatment that is correlated with the treatment effect. What if instead site DMs have negligible information about the site-specific treatment effect, and selection into treatment is instead driven by other site-specific factors uncorrelated with treatment? It is an open question which scenario is more common empirically. In this Section, we nest the two scenarios and examine how the bias comparison is affected.

Under Lemma 1 conditions, the baseline model predicts selection into treatment at site s iff $\tau_s^O > -\overline{\tau}$. We define an alternative selection variable, χ_s^O, observed by the site-s DM but not the researcher, and distributed across sites identically to τ_s^O, but independently of τ_s^O and all components of $\overline{\tau}_s$: $\{\nu_{sj}^O\}$, $\{\nu_{sj}^U\}$, and u_{ij}. In this Section's modified model, a fraction of sites $\gamma \in [0, 1]$ are assumed to be "informed" and thus select into treatment – autonomous or experimental – iff $\tau_s^O > -\overline{\tau}$. The remaining $1-\gamma$ "uninformed" sites are assumed to select into treatment – autonomous or experimental – iff $\chi_s^O > -\overline{\tau}$. Thus, uninformed sites select into treatment at the same rate as informed sites, but do so for reasons that are orthogonal to the treatment effect (satisfying the condition for external validity in Hotz et al., 2005, and Alcott, 2015). Being informed is independent of all other site variables.

The bias in the experimental case can now be written (modifying Equation 3) as

$$\begin{aligned}\gamma E\left(\tau_i \mid \tau_{S_i}^O > -\overline{\tau}\right) + (1-\gamma) E\left(\tau_i \mid \chi_{S_i}^O > -\overline{\tau}\right) - \overline{\tau} \\ = \gamma \left[E\left(\nu_{s1}^O \mid \tau_s^O > -\overline{\tau}\right) - E\left(\nu_{s0}^O \mid \tau_s^O > -\overline{\tau}\right) \right] ,\end{aligned}$$

where the equality follows from the independence of χ_s^O from $\overline{\tau}_s$. The less informed site DMs are (smaller γ), the less bias the experimental approach contains, because the experiment is closer to randomly assigned across sites with respect to the treatment effect. If sites are completely uninformed ($\gamma = 0$), there is no bias.

The bias in the observational case is now (modifying Equation 1)

$$\begin{aligned}\gamma \ \left[E\left(\tau_i \mid \tau_{S_i}^O > -\overline{\tau}\right) - \overline{\tau} + E\left(Y_{i0} \mid \tau_{S_i}^O > -\overline{\tau}\right) - E\left(Y_{i0} \mid \tau_{S_i}^O \leq -\overline{\tau}\right)\right] \\ (1-\gamma) \ \left[E\left(\tau_i \mid \chi_{S_i}^O > -\overline{\tau}\right) - \overline{\tau} + E\left(Y_{i0} \mid \chi_{S_i}^O > -\overline{\tau}\right) - E\left(Y_{i0} \mid \chi_{S_i}^O \leq -\overline{\tau}\right)\right] \\ = \gamma \left[E\left(\nu_{s1}^O \mid \tau_s^O > -\overline{\tau}\right) - E\left(\nu_{s0}^O \mid \tau_s^O \leq -\overline{\tau}\right) \right] ,\end{aligned}$$

where again the equality follows from the independence of χ_s^O from all components of $\overline{\tau}_s$. Here too, a greater prevalence of uninformed sites (smaller γ) leads to smaller bias, again because treatment is closer to randomly assigned, and there is no bias if sites are completely uninformed.

Remarkably, the bias comparison between the two approaches is unaffected by this lack of information: both biases are simply scaled toward zero by the same factor γ (compare the above biases to those in Equations 9 and 10). In this context, the experimental approach gains external validity the more that sites select into treatment for reasons orthogonal to their site-specific treatment effect. But at the same rate, the internal validity of the observational approach increases. This is no coincidence; the bias is similar in both cases, though it manifests differently. In short, the more lacking in information sites are, the better an experiment performs but the more unnecessary it is for identification.

A different conclusion may be reached under an alternative assumption that χ_s^O is un-

correlated with the treatment effect $\overline{\tau}_s$, but correlated with baseline outcomes at site s, $E(Y_{i0}|S_i = s)$. Under this alternative assumption, the experimental bias continues to decline toward zero in γ while the observational bias need not approach zero in γ, since $E(Y_{i0}|\chi^O_{S_i} > -\overline{\tau}) - E(Y_{i0}|\chi^O_{S_i} \leq -\overline{\tau})$ may be non-zero. If so, the experimental approach could be less biased if sites were sufficiently uninformed. However, it is not clear how natural this assumption would be, as it involves χ^O_s correlated with ν^O_{s0} or ν^U_{s0} while being uncorrelated with $\nu^O_{s1} - \nu^O_{s0}$ and $\nu^U_{s1} - \nu^U_{s0}$.

At any rate, this Section demonstrates that even a significant lack of information across sites may do nothing to alter the bias comparison between the experimental and the simple observational approach.

5.2 Biased expectations

In the baseline model, site DMs have unbiased beliefs about the site-specific and population ATE. Here we allow for site DMs to have biased beliefs.

We modify the baseline model by assuming that all site DMs make decisions based on prior beliefs about $\{\nu^U_{sj}\}$ that induce a prior distribution $\mathcal{F}$ over τ^U_s. Letting $E^{\mathcal{F}}$ denote the subjective expectations based on prior $\mathcal{F}$, we assume $E^{\mathcal{F}}(\tau^U_s) = -\pi$, and that τ^U_s is independent of $\{\nu^O_{sj}\}$ under $\mathcal{F}$. The remainder of the model is unchanged, including that subjective expectations of all other random variables correspond to reality and that $E(\tau^U_s) = 0$. Note that π measures (unfounded) pessimism; the site-s DM expected site-specific average treatment effect is

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

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

Now autonomous selection into treatment occurs iff $\overline{\tau}^{DM,\mathcal{F}}_s > 0$, i.e. $T_i = 1$ iff $S_i \in \mathcal{A}_1^{-\pi} \equiv$

$\{s \in \mathcal{S} : \tau_s^O > -\overline{\tau} + \pi\}$. Selection into experimental treatment occurs iff

$$(1-\delta)\lambda\left(\overline{\tau}+\tau_s^O-\pi\right)+\delta(1-\phi)\,max\left\{\overline{\tau}+\tau_s^O-\pi,0\right\}+\delta\phi\,E^{\mathcal{F}}\left(max\left\{\overline{\tau}+\tau_s^O+\tau_s^U,0\right\}\right) > (1-\delta)k\,,$$

modifying Inequality 8. An analogous version of Lemma 1 continues to hold and provide conditions under which selection into treatment occurs at exactly the same sites in the experimental case as in the autonomous case. Under those conditions, we can write the total bias in the observational case as

$$E\left(\nu_{s1}^O \mid \tau_s^O > -\overline{\tau} + \pi\right) - E\left(\nu_{s0}^O \mid \tau_s^O \leq -\overline{\tau} + \pi\right) \tag{12}$$

and in the experimental case as

$$E\left(\nu_{s1}^O \mid \tau_s^O > -\overline{\tau} + \pi\right) - E\left(\nu_{s0}^O \mid \tau_s^O > -\overline{\tau} + \pi\right)\ , \tag{13}$$

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 estimates of treated outcomes are based – are a more selected sample the more pessimistic site DMs are. 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 about treatment would seem to tilt the bias comparison in favor of the observational case, since its untreated outcomes are estimated from opt-outs, which are more representative under more pervasive pessimism. More precisely and analogous to Proposition 1, we have the following result.

Proposition 2. *Under Assumption A2, and assuming identical selection in the observational and experimental approaches, i.e.* $\mathcal{S}_1 = \mathcal{S}_1^E = \mathcal{A}_1^{-\pi} \equiv \{s \in \mathcal{S} : \tau_s^O > -\overline{\tau} + \pi\}$*, for every* $\overline{\tau}$

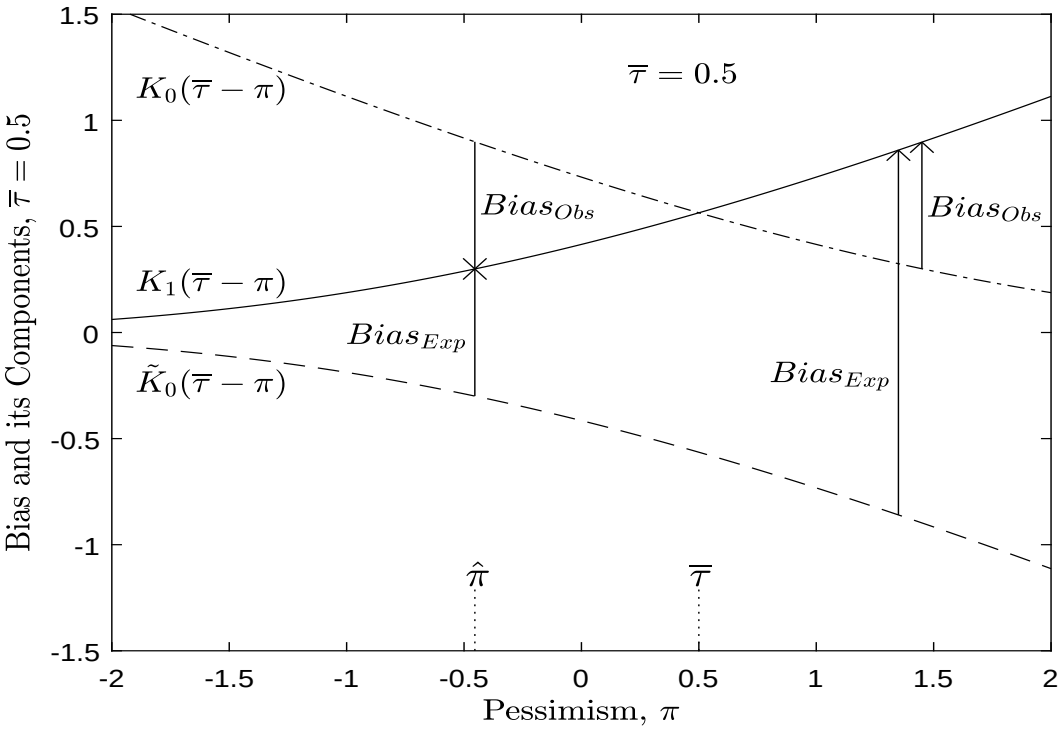


Figure 3: Displays bias of the observational and experimental approaches for different values of the degree of pessimism π, given the average treatment effect satisfies $\overline{\tau} = 0.5$. Note that $Bias_{Obs} = K_1(\overline{\tau} - \pi) - K_0(\overline{\tau} - \pi)$ and $Bias_{Exp} = K_1(\overline{\tau} - \pi) - \tilde{K}_0(\overline{\tau} - \pi)$. Here ν_{s0}^O and ν_{s1}^O are iid standard normal, and the observational approach is less biased iff $\pi > \hat{\pi} = \overline{\tau} - \hat{\tau} = 0.5 - \sqrt{2}\,\Phi^{-1}(3/4) \approx -0.45$.

there exists a $\hat{\pi}$, such that the observational bias is smaller in magnitude if $\pi > \hat{\pi}$ and the experimental bias is smaller in magnitude if $\pi < \hat{\pi}$. Assuming further that ν_{s0}^O and ν_{s1}^O are iid, then $\hat{\pi} < \overline{\tau}$.

That is, sufficient pessimism about the program's treatment effect (π) means that the observational approach is less biased. If the site-specific errors ν_{s0}^O and ν_{s1}^O are iid, the cutoff level of pessimism is less than the ATE; for example, if the ATE is zero, the observational approach is less biased if site DMs are pessimistic ($\pi > 0$) or moderately optimistic ($\pi \in (\hat{\pi}, 0)$).

Figure 3 illustrates the Proposition's logic. More precisely, using the notation from Equations 11-13, we have that $Bias_{Obs} = K_1(\overline{\tau} - \pi) - K_0(\overline{\tau} - \pi)$ and $Bias_{Exp} = K_1(\overline{\tau} - \pi) - \tilde{K}_0(\overline{\tau} - \pi)$; the Figure graphs these components for $\{\nu_{sj}^O\}$ iid standard normal and $\overline{\tau} = 0.5$. The observational and experimental biases are increasing in π; the experimental

bias increases from zero, while the observational bias increases through zero and is always smaller (closer to $-\infty$). Thus at some cutoff value for π, here $\hat{\pi} \approx -0.45$, the observational bias is as negative as the experimental bias is positive; to the right (left), the observational (experimental) bias is smaller in magnitude.

Thus, greater (unfounded) pessimism among site DMs tilts the comparison in favor of the observational approach, and the reverse for optimism. The reason is that greater pessimism (optimism) means that sites opting out (in) become more representative of the population, and these are the sites differentially used by the observational (experimental) approach.

In sum, the paper's argument does not depend on site DMs correctly knowing the ATE. Further, the degree of pessimism about the treatment is another factor that can tilt the comparison of biases from the experimental and observational approaches.

5.3 Autonomous selection with 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 autonomously adopt it for some period of time, and that they may discard it after learning about it.

We assume that if sites can learn autonomously, 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 are weighed along with expected long-run payoffs after learning and then choosing optimally. The decision is then governed by Inequality 7 of the experimental case, except with potentially different parameters (superscripted by "A" for autonomous) capturing the fraction treated in the learning phase, λ^A, the probability of successfully learning, ϕ^A, discounting of the learning phase (which depends in part on how long that phase is), δ^A, and the cost of learning, k^A. Simplification leads to a condition

identical to Inequality 8 for whether a site selects treatment:

$$(1-\delta^A)\lambda^A\left(\overline{\tau}+\tau_s^O\right)+\delta^A(1-\phi^A)\,max\left\{\overline{\tau}+\tau_s^O,0\right\}+\delta^A\phi^A\,E\left(max\left\{\overline{\tau}+\tau_s^O+\tau_s^U,0\right\}\left|\tau_s^O\right.\right) > (1-\delta)k^A\ .$$

There are now two straightforward ways for site selection into initial treatment to be the same in experimental and autonomous cases. It is identical if all relevant parameters are the same, i.e. $\lambda=\lambda^A$, $\phi=\phi^A$, $\delta=\delta^A$, and $k=k^A$. Even if parameters differ, it is also identical (at least asymptotically) if Lemma 1 conditions hold for both experimental and autonomous parameters. In this Section, we focus on the latter case and assume that selection into initial treatment is identical and captured by $\mathcal{S}_1=\mathcal{S}_1^E=\mathcal{A}_1\equiv\{s\in\mathcal{S}:\tau_s^O>-\overline{\tau}\}$.

However, there is a further stage of selection that may be relevant in the autonomous case, as some sites that opt into treatment for the trial period may opt out of treatment after learning but before the researcher measures outcomes. This is not an issue in the experimental approach, where by design the researcher measures outcomes while treatment is in place; whether it is an issue with the observational approach depends on whether outcomes are measured during the early phase while learning is happening, or during the mature phase when any learning is complete. We thus discuss two cases differing in the timing of the study relative to autonomous learning.

I) Consider first *early-phase study* and identical initial selection due to Lemma 1 conditions holding. In this case, selection is the same as that analyzed in Sections 4.1 and 4.2, and 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 determined predominantly by initial estimates of site-specific treatment effects in both cases.

II) Next, consider *mature-phase study* and identical initial selection due to Lemma 1 conditions holding. In this case, the observational researcher is observing $T_i=1$ only for sites that initially opted in *and* that did not later opt out after learning. Learning is probabilistic, happening with probability ϕ^A at sites adopting the treatment. Let $L_s^A=1$ iff the site-s

DM would learn $\{\nu^U_{sj}\}$ after adopting (unobserved by the site DM); so $Prob(L^A_s = 1) = \phi^A$. Then the researcher observes $T_i = 1$ iff $S_i \in \mathcal{A}'_1$, where

$$\mathcal{A}'_1 \equiv \left\{ s \in \mathcal{S} : \left(L^A_s = 0 \;\;\&\;\; \tau^O_s > -\overline{\tau}\right) \;\mid\; \left(L^A_s = 1 \;\;\&\;\; \tau^O_s + \min\{0, \tau^U_s\} > -\overline{\tau}\right) \right\}. \tag{14}$$

This is because to be observed in treatment, sites must have opted in initially ($\tau^O_s > -\overline{\tau}$) and, if they have then learned, also have a positive true sATE ($\tau^O_s + \tau^U_s > -\overline{\tau}$). Modifying equations 5 and 6, the total bias in the observational estimation can be written

$$Bias_{Obs} = Bias\text{-}1 + Bias\text{-}2 = E\left(\nu^O_{s1} + \nu^U_{s1} \mid s \in \mathcal{A}'_1\right) - E\left(\nu^O_{s0} + \nu^U_{s0} \mid s \in \mathcal{A}'_0\right), \tag{15}$$

where $\mathcal{A}'_0 \equiv \mathcal{S} \setminus \mathcal{A}'_1$.

Similar to Section 4.3, assume that ν^O_{s1}, ν^O_{s0}, ν^U_{s1}, and ν^U_{s0} are distributed with no mass points over support $\mathbb{R}$. For $x \in \mathbb{R}$, define

$$\begin{aligned} K_{1M}(x) \equiv & E\left(\nu^M_{s1} \mid \tau^O_s + \min\{0, \tau^U_s\} > -x\right), \qquad M \in \{O, U\} \\ K_{0M}(x) \equiv & E\left(\nu^M_{s0} \mid \tau^O_s + \min\{0, \tau^U_s\} \leq -x\right), \qquad M \in \{O, U\}\ . \end{aligned} \tag{16}$$

Using notation $K_1(\overline{\tau})$ and $K_0(\overline{\tau})$ defined for Equation 11, we show that

Lemma 2. *If observational selection is defined by $\mathcal{S}_1 = \mathcal{A}'_1$ defined in Equation 14, then* $Bias_{Obs} = (1 - \phi^A)\left[K_1(\overline{\tau}) - K_0(\overline{\tau})\right] + \phi^A\left[K_{1O}(\overline{\tau}) - K_{0O}(\overline{\tau}) + K_{1U}(\overline{\tau}) - K_{0U}(\overline{\tau})\right]$.

Unsurprisingly, the bias here is a convex combination of the bias without learning, the first bracketed term (analyzed in Section 4), and the bias with learning, the second bracketed term. This learning-bias involves potentially biased estimates of both the $\{\nu^O_{sj}\}$ and the $\{\nu^U_{sj}\}$, since some sites have learned the latter before making their final selection decision.

Note that

$$\lim_{x\to\infty} K_{1O}(x) = \lim_{x\to-\infty} K_{0O}(x) = \lim_{x\to\infty} K_{1U}(x) = \lim_{x\to-\infty} K_{0U}(x) = 0\ ,$$

since ν_{s1}^O, ν_{s0}^O, ν_{s1}^U, and ν_{s0}^U are mean-zero. Since $K_{1O}(x) = E(\nu_{s1}^O | \nu_{s1}^O > \nu_{s0}^O - x - \min\{0, \tau_s^U\})$, it is reasonable to assume that $K'_{1O}(x) < 0$; extending this logic supports the assumption that

$$K'_{1O}(x) < 0, \qquad K'_{0O}(x) > 0, \qquad K'_{1U}(x) < 0, \qquad K'_{0U}(x) > 0 \ . \tag{A3}$$

The following result can be shown.

Proposition 3. *Under Assumptions A2 and A3, and assuming identical* initial *selection in the observational and experimental approaches and* mature-phase *study in the observational case, so that selection in the experimental case is described by* $\mathcal{S}_1^E = \mathcal{A}_1 \equiv \{s \in \mathcal{S} : \tau_s^O > -\overline{\tau}\}$ *and in the observational case by* $\mathcal{S}_1 = \mathcal{A}'_1$ *of Equation 14, there exist* $\hat{\tau}_{lo} \in \Re \cup \{-\infty\}$ *and* $\hat{\tau} > \hat{\tau}_{lo}$*, such that the observational bias is smaller in magnitude if* $\overline{\tau} \in (\hat{\tau}_{lo}, \hat{\tau})$ *and the experimental bias is smaller in magnitude if* $\overline{\tau} > \hat{\tau}$*.*

Thus, very similarly to the baseline case, the observational approach delivers less bias if $\overline{\tau}$ is not too large, and perhaps not too small either. We illustrate the Proposition's logic in Figure 4. It graphs $K_1(\overline{\tau})$, $K_0(\overline{\tau})$, and $\tilde{K}_0(\overline{\tau})$ with dotted lines, $K_{1O}(\overline{\tau})$ and $K_{0O}(\overline{\tau})$ with solid lines, and $K_{1U}(\overline{\tau})$ and $K_{0U}(\overline{\tau})$ with dashed lines for $\{\nu_{sj}^O\}$ and $\{\nu_{sj}^U\}$ iid standard normal. The observational and experimental biases come from the differences between the relevant functions; since the observational bias is a set of weighted differences, a separate graph is used to picture the biases, Figure 5.

The experimental bias is graphed with a dash-dotted line, while the two components of the observational bias are graphed with solid and dashed lines, corresponding to all sites learning after adopting ($\phi^A = 1$) and no sites learning ($\phi^A = 0$), respectively; the true observational bias is a convex combination of these determined by ϕ^A. As before, both biases are decreasing in $\overline{\tau}$, the experimental bias toward zero and the observational bias through zero. Thus at some cutoff value for $\overline{\tau}$, the observational bias is as negative as the experimental bias is positive; to the left (right), the observational (experimental) bias is smaller in magnitude. We cannot rule out in general that the observational bias exceeds the

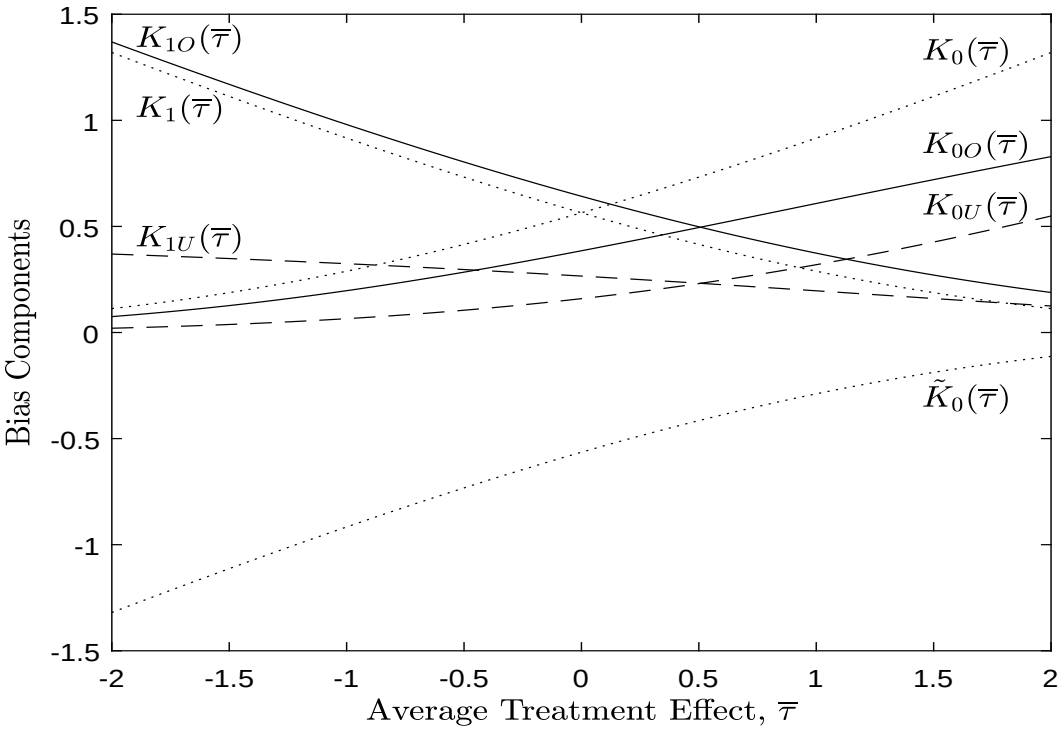


Figure 4: Displays the components of the bias of the observational and experimental approaches for different values of the average treatment effect, $\overline{\tau}$. Note that $Bias_{Obs} = (1-\phi^A)\left[K_1(\overline{\tau}) - K_0(\overline{\tau})\right] + \phi^A\left[K_{1O}(\overline{\tau}) - K_{0O}(\overline{\tau}) + K_{1U}(\overline{\tau}) - K_{0U}(\overline{\tau})\right]$ (see Lemma 2) and $Bias_{Exp} = K_1(\overline{\tau}) - \tilde{K}_0(\overline{\tau})$. Here ν_{s0}^O, ν_{s1}^O, ν_{s1}^U, and ν_{s0}^U are iid standard normal.

experimental bias for $\overline{\tau}$ low enough too, though we have not observed this in computation.

The case of no learning ($\phi^A = 0$) is analyzed in Section 4.3, so consider the novel case of perfect learning ($\phi^A = 1$). In this computational example, learning extends the space over which the observational approach has less bias – one sees this since $\hat{\tau}_{\phi^A=0} < \hat{\tau}_{\phi^A=1} \approx 1.19$. This is a bit surprising as learning tends to make the observational bias more positive: sites are more acutely selected because they have responded to experiential learning about the site-specific treatment effect. Indeed, note from Figure 5 that the observational bias is higher when $\phi^A = 1$ than when $\phi^A = 0$. But, this implies that the magnitude (absolute value) of the observational bias with learning is smaller over some range, and in particular over the key range where the two approaches give similar bias magnitudes. Learning raises the observational bias, but since the observational bias must be negative to be larger in

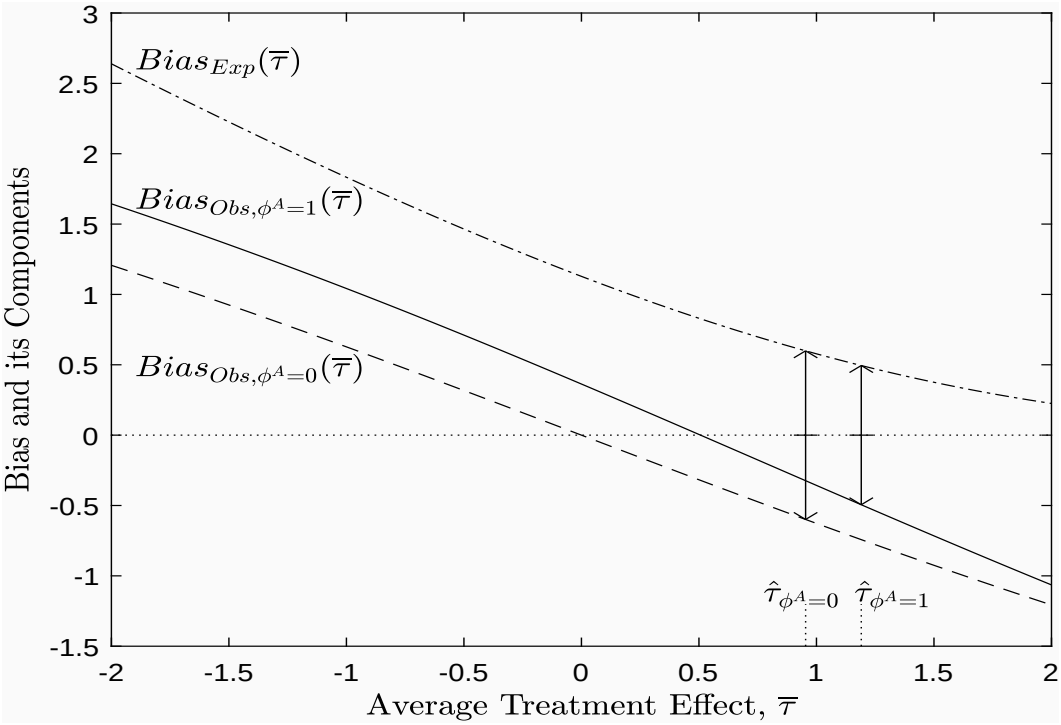


Figure 5: Displays the bias of the experimental and observational approaches, the latter with full learning ($\phi^A = 1$, solid line) and no learning ($\phi^A = 0$, dashed line), for different values of the average treatment effect, $\overline{\tau}$. Here ν^O_{s0}, ν^O_{s1}, ν^U_{s1}, and ν^U_{s0} are iid standard normal. The experimental approach is less biased over this range than the observational approach with no learning iff $\overline{\tau} > \hat{\tau}_{\phi^A=0} = \sqrt{2}\,\Phi^{-1}(3/4) \approx 0.95$ and the observational approach with learning iff $\overline{\tau} > \hat{\tau}_{\phi^A=1} \approx 1.19$.

magnitude, this pushes out the region for which the observational bias remains smaller in magnitude.

On the whole, then, autonomous learning can exacerbate the observational bias but may still make the observational approach less biased over more of the parameter space. More generally, this complication of the model does not obviously push the comparison in one direction or another.[33]

[33]We have ignored autonomous learning as an option in the experimental case. If instead sites approached by an experimenter also have the option to learn about the treatment and test it themselves, selection into the experiment would involve similar complications to the ones introduced here.

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. Further, such site selection bias can lead internally valid RCT estimates to exceed simple observational estimates in bias for the ATE.

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. Given that identification is for 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 standard RCT practice to augment substantially general knowledge about high-stakes treatments may be somewhat 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. Ideally, the site selection process should be documented, including identity of all sites approached and negotations with partner sites that determined the specific policy tested, so the reader can gauge the nature of selection into the experiment.[34]

Estimating the ATE may not always be the most fruitful approach. A growing literature explores optimal estimation targets and treatment assignment, in some cases incorporating

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

self-selection into treatment; see, e.g., Kitagawa and Tetenov (2018) and Ida et al. (2023). We view this literature as complementary to the current study; both suggest drawbacks to simple experimental estimates of the ATE and their application. However, there continues to be a large body of literature focused on estimating simple average treatment effects experimentally, and this paper is aimed at interpreting this literature.

We do 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 treatments (e.g. with the level of risk of the treatment). 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, degree of pessimism, and so on – can provide a basis for identification assumptions in the RCT context.

The analysis does point to 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; this eliminates the selection problem. 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. A potential shortcoming of this approach is that in some contexts, such a liquidity infusion could affect treated or untreated outcomes. An alternative to lump-sum compensation is to insure the partner site against the downside risk of the treatment; this seems promising, though it may raise the concern of distorted site-level incentives, i.e. moral hazard.

Another solution is to adapt the approaches used to deal with individual-level selection into treatment – i.e. estimating ITT or LATE parameters – to address site-level selection. Identifying a site-level ITT would be possible after defining a set of potential partner sites,

extending experimental treatment offers to a random subset of these sites, and tracking outcomes at the entire set of potential sites. Assuming no spillovers, the related LATE could also be identified. However, the LATE would be an ATE for "compliers", i.e. willing sites that select into experimental treatment – raising the familiar concern from our model that the same "Bias–1" can plague both the observational and experimental cases (relative to the ATE). Unwilling ("never-taking") sites would be on standard ground treating the evidence as "non-causal".

An adaptation of Brinch et al. (2017) would enable a test for external validity of the experimental approach. It would not require data from non-participating sites, but would require data from multiple participating sites and a possibly binary instrument that shifts willingness to adopt experimental treatment. The idea is that as the instrument shifts the site selection margin, selection bias should be detectible in differences in treatment effect magnitudes at different margins. As a simple example, one could recruit sites randomly assigning a site-level bonus payment or no bonus. Assuming the bonus changes the probability of assenting to the experiment (instrument relevance), a comparison of the bonus and no-bonus sATEs would constitute a test for external validity.[35]

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

A promising approach from beyond the model is to use observables in both selection into treatment and as factors behind the treatment effect to model or control for site selection. There is a growing body of work in this direction, as noted in Section 2. It is not guaranteed to solve the problem of site selection on unobservables, however (Allcott, 2015). And if external validity biases can easily swamp the internal validity biases of a naive observational

[35]Gechter and Meager (2022) demonstrate the significant possibilities of this kind of approach. They differ from our work in using a different definition of site selection, based on researcher's choice of whether to use an experimental or observational technique to estimate the sATE, rather than site manager's decision to assent to experimental treatment. They also do not examine settings piloting novel, "risky" treatments.

approach in a setting without observables, they could plausibly continue to do so with observables employed to ameliorate biases in both settings.

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?[36] In some settings, a case may be made that the ATE for willing sites is the desired parameter, since these sites may be the ones open to consider adopting. If so, this has strong implications for how the research findings should be publicized: not as a program that is expected to work for an arbitrary site, but specifically for a site to whom the program initially sounds interesting and feasible.

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 theory suggests that simply assuming orthogonality of selection into risky experimentation is questionable, and that the further assumptions required to be confident in a smaller experimental bias seem substantive and worth making explicit.

[36]Specifically, for a given experiment, if the treatment was instead autonomously selected across a number of sites and a cross-site observational study carried out, would such a study be vulnerable to the (internal validity) critique of site selection bias? If so, is it reasonable to assume a similar site selection bias is not similarly biasing results from an RCT, 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 Banerjee, Dean Karlan, and Jonathan Zinman. Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, 7(1):1–21, January 2015.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] Christian N. Brinch, Magne Mogstad, and Matthew Wiswall. Beyond LATE with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039, August 2017.

[11] 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.

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

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

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Takanori Ida, Takunori Ishihara, Koichiro Ito, Daido Kido, Toru Kitagawa, Shosei Sakaguchi, and Shusaku Sasaki. Choosing who chooses: Selection-driven targeting in energy rebate programs. NBER Working Paper, September, 2022.

[25] 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.

[26] Toru Kitagawa and Aleksey Tetenov. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, March 2018.

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

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

[29] 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.

[30] 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.

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

[32] 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.

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

[34] 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.

[35] 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.

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

[37] Bruce Wydick. Microfinance on the margin: Why recent impact studies may understate average treatment effects. *Journal of Development Effectiveness*, 8(2):257–265, 2016.

A Proofs

Proof of Lemma 1A. Assent to the experiment occurs iff $s \in \mathcal{S}_1^E$, where $\mathcal{S}_1^E \equiv \{s \in \mathcal{S} :$ Inequality 8 holds for $s\}$. Given that Inequality 8's second and third terms are weakly positive and that $k = 0$, the inequality clearly holds if $\overline{\tau} + \tau_s^O > 0$, i.e. for all $s \in \mathcal{A}_1$; thus $\mathcal{A}_1 \subseteq \mathcal{S}_1^E$. Fix $\epsilon > 0$. It remains to show that under conditions a), b), or c), if $s \notin \mathcal{A}_1^{+\epsilon}$ then $s \notin \mathcal{S}_1^E$, i.e. that if $\overline{\tau} + \tau_s^O \leq -\epsilon$, then Inequality 8 is violated. If $\overline{\tau} + \tau_s^O \leq -\epsilon$, the negation of Inequality 8 with $k = 0$ is equivalent to

$$\overline{\tau} + \tau_s^O \leq -\frac{\delta\phi\, E\left(max\left\{\overline{\tau} + \tau_s^O + \tau_s^U, 0\right\}\right)}{\lambda(1-\delta)} \equiv -Q \ ,$$

say. The expectation term is bounded by 0 and B^U (using A1 and that $\overline{\tau} + \tau_s^O \leq -\epsilon$). Clearly if δ or ϕ is small enough (e.g. $\phi \leq \epsilon\lambda(1-\delta)/(\delta B^U)$), then $Q \leq \epsilon$; the same holds if $B^U \leq \epsilon\lambda(1-\delta)/(\delta\phi)$. It follows that under any of the three conditions, $\overline{\tau} + \tau_s^O \leq -\epsilon$ guarantees that $\overline{\tau} + \tau_s^O \leq -Q$. ■

Proof of Lemma 1B. Assent to the experiment occurs iff Inequality 8 holds. Note that its left-hand side is strictly increasing in $\overline{\tau} + \tau_s^O$. Further, if $\overline{\tau} + \tau_s^O = 0$ and $k = \hat{K}$, then its left-hand side and right-hand side are equal. Thus, for $k = \hat{K}$, inequality 8 is satisfied at site s iff $\overline{\tau} + \tau_s^O > 0$. It also follows that if $k > \hat{K}$, a stricter condition on $\overline{\tau} + \tau_s^O$ is necessary and sufficient, while if $k < \hat{K}$, a weaker condition is. ■

Proof of Proposition 1. Note that $Bias_{Obs} < Bias_{Exp}$ for any $\overline{\tau}$. This follows since if $\mathcal{S}_1 = \mathcal{S}_1^E = \mathcal{A}_1$, $Bias_{Obs} = K_1(\overline{\tau}) - K_0(\overline{\tau})$ and $Bias_{Exp} = K_1(\overline{\tau}) - \tilde{K}_0(\overline{\tau})$; and further, $K_0(\overline{\tau}) > \tilde{K}_0(\overline{\tau})$, since both $K_0(\overline{\tau})$ and $\tilde{K}_0(\overline{\tau})$ are strictly increasing by Assumption A2, one from zero and the other toward zero (since $\lim_{x\to-\infty} K_0(x) = \lim_{x\to\infty} \tilde{K}_0(x) = 0$). Also, both biases are strictly decreasing in $\overline{\tau}$ (since $K_1'(x) < 0$ while $K_0'(x) > 0$ and $\tilde{K}_0'(x) > 0$, by Assumption A2). The experimental bias is thus also strictly positive (since $\lim_{x\to\infty}[K_1(x) - \tilde{K}_0(x)] = 0$). These facts imply that the observational bias can only be larger if it is negative; and further that necessary and sufficient for $|Bias_{Obs}| > |Bias_{Exp}|$ is that $-Bias_{Obs} > Bias_{Exp}$, i.e. that $K_0(\overline{\tau}) + \tilde{K}_0(\overline{\tau}) > 2K_1(\overline{\tau})$. The cutoff $\hat{\tau}$ then follows from the intermediate value theorem using the continuity (following from no mass points) and monotonicity of all functions in this inequality and the limits $\lim_{x\to-\infty} \tilde{K}_0(x) < 0 = \lim_{x\to-\infty} K_0(x) < \lim_{x\to-\infty} K_1(x)$ and $\lim_{x\to\infty} \tilde{K}_0(x) = \lim_{x\to\infty} K_1(x) = 0 < \lim_{x\to\infty} K_0(x)$.

If ν_{s0}^O and ν_{s1}^O are iid, then $K_1(0) = K_0(0)$, so $Bias_{Obs}$ is weakly positive for $\overline{\tau} \leq 0$. Since the observational bias can only be larger if it is strictly negative, as argued in the previous

paragraph, the cutoff $\hat{\tau}$ must be strictly positive. ■

Proof of Corollary 1. By Proposition 1, there exists a $\hat{\tau}$ such that the observational (experimental) bias is smaller if $\overline{\tau} < \hat{\tau}$ ($\overline{\tau} > \hat{\tau}$). Let $\hat{p} \equiv p(\hat{\tau})$. It follows from the fact that $p(\overline{\tau})$ is monotonically increasing in $\overline{\tau}$ that the observational (experimental) bias is smaller if $p(\overline{\tau}) < \hat{p}$ ($p(\overline{\tau}) > \hat{p}$). ■

Proof of Proposition 2. Fix $\overline{\tau}$. Note that $Bias_{Obs} < Bias_{Exp}$ for any π. This follows since if $\mathcal{S}_1 = \mathcal{S}_1^E = \mathcal{A}_1^{-\pi}$, $Bias_{Obs} = K_1(\overline{\tau} - \pi) - K_0(\overline{\tau} - \pi)$ and $Bias_{Exp} = K_1(\overline{\tau} - \pi) - \tilde{K}_0(\overline{\tau} - \pi)$; and further, $K_0(\overline{\tau} - \pi) > \tilde{K}_0(\overline{\tau} - \pi)$, as established in the Proof of Proposition 1. Also, both biases are strictly increasing in π (since $K_1'(x) < 0$ while $K_0'(x) > 0$ and $\tilde{K}_0'(x) > 0$, by Assumption A2). The experimental bias is also strictly positive, as established in the Proof of Proposition 1. These facts imply that the observational bias can only be larger if it is negative; and further that necessary and sufficient for $|Bias_{Obs}| > |Bias_{Exp}|$ is that $-Bias_{Obs} > Bias_{Exp}$, i.e. that $K_0(\overline{\tau}-\pi)+\tilde{K}_0(\overline{\tau}-\pi) > 2K_1(\overline{\tau}-\pi)$. The cutoff $\hat{\pi}$ then follows from the intermediate value theorem using the continuity (following from no mass points) and monotonicity of all functions in this inequality and the limits $\lim_{\pi\to-\infty} \tilde{K}_0(\overline{\tau} - \pi) = \lim_{\pi\to-\infty} K_1(\overline{\tau} - \pi) = 0 < \lim_{\pi\to-\infty} K_0(\overline{\tau} - \pi)$ and $\lim_{\pi\to\infty} \tilde{K}_0(\overline{\tau} - \pi) < 0 = \lim_{\pi\to\infty} K_0(\overline{\tau} - \pi) < \lim_{\pi\to\infty} K_1(\overline{\tau} - \pi)$.

If ν_{s0}^O and ν_{s1}^O are iid, then $K_1(0) = K_0(0)$, so $Bias_{Obs}$ is weakly positive for $\pi \geq \overline{\tau}$. Since the observational bias can only be larger if it is strictly negative, the cutoff $\hat{\pi}$ must satisfy $\hat{\pi} < \overline{\tau}$. ■

Proof of Lemma 2. Note that

$$
\begin{aligned}
&E(\nu_{s1}^O | s \in \mathcal{A}_1') \\
&\quad = E(\nu_{s1}^O | L_s^A = 1, s \in \mathcal{A}_1') \times Prob(L_s^A = 1) + E(\nu_{s1}^O | L_s^A = 0, s \in \mathcal{A}_1') \times Prob(L_s^A = 0) \\
&\quad = \phi^A \, E(\nu_{s1}^O | \tau_s^O + \min\{0, \tau_s^U\} > -\overline{\tau}) + (1 - \phi^A) E(\nu_{s1}^O | \tau_s^O > -\overline{\tau}) \\
&\quad = \phi^A K_{1O}(\overline{\tau}) + (1 - \phi^A) K_1(\overline{\tau}) \; .
\end{aligned}
$$

Similarly,

$$
E(\nu_{s0}^O | s \in \mathcal{A}_0') = \phi^A \, E(\nu_{s0}^O | \tau_s^O + \min\{0, \tau_s^U\} \leq -\overline{\tau}) + (1 - \phi^A) E(\nu_{s0}^O | \tau_s^O \leq -\overline{\tau}) = \phi^A K_{0O}(\overline{\tau}) + (1 - \phi^A) K_0(\overline{\tau}) \; .
$$

Now

$$E(\nu_{s1}^U | s \in \mathcal{A}_1') = \phi^A \, E(\nu_{s1}^U | \tau_s^O + \min\{0, \tau_s^U\} > -\overline{\tau}) + (1-\phi^A) E(\nu_{s1}^U | \tau_s^O > -\overline{\tau}) = \phi^A K_{1U}(\overline{\tau}) \ ,$$

the first equality by above reasoning and the last because ν_{s1}^U is mean-zero and independent of $\{\nu_{sj}^O\}$. Similarly, $E(\nu_{s0}^U | s \in \mathcal{A}_0') = \phi^A K_{0U}(\overline{\tau})$. Using these expressions in Equation 15 gives the result. ■

Proof of Proposition 3. Note that both biases are strictly decreasing in $\overline{\tau}$, by Assumptions A2 and A3 (under the assumptions on selection, $Bias_{Obs}$ is given in Lemma 2 and $Bias_{Exp} = K_1(\overline{\tau}) - \tilde{K}_0(\overline{\tau})$). The experimental bias declines toward zero (since $\lim_{x\to\infty} K_1(x) = \lim_{x\to\infty} \tilde{K}_0(x) = 0$), while the observational bias declines through zero (since $\lim_{x\to\infty} K_1(x) = \lim_{x\to\infty} K_{1O}(x) = \lim_{x\to\infty} K_{1U}(x) = 0$, while $K_0(x)$, $K_{0O}(x)$, and $K_{0U}(x)$ are all strictly increasing and thus, given their zero-limits as $x \to -\infty$, strictly positive and bounded away from zero for $x \geq 0$). This implies that $Bias_{Obs}(x) < Bias_{Exp}(x)$ for x large enough; but it may be that $Bias_{Obs}(x) \geq Bias_{Exp}(x)$ for x small enough. Define $\hat{\tau}_{lo}$ as the maximum value for x such that $Bias_{Obs}(x) \geq Bias_{Exp}(x)$, or $-\infty$ if $Bias_{Obs}(x) < Bias_{Exp}(x)$ for all x. Now for $x > \hat{\tau}_{lo}$, $Bias_{Obs}(x) < Bias_{Exp}(x)$, so the observational bias can be larger *in magnitude* only if it is negative and $-Bias_{Obs}(x) > Bias_{Exp}(x)$. Since the LHS of this inequality is strictly increasing in x beyond zero, and the RHS is strictly decreasing in x toward zero, the observational bias is larger for all $x > \hat{\tau}$, where $\hat{\tau}$ is implicitly defined by $-Bias_{Obs}(\hat{\tau}) = Bias_{Exp}(\hat{\tau})$. For $x \in (\hat{\tau}_{lo}, \hat{\tau})$, the observational bias is smaller; it is either positive, so that the magnitude comparison is guaranteed by $Bias_{Obs}(x) < Bias_{Exp}(x)$, or it is negative, in which case the magnitude comparison is given by $-Bias_{Obs}(x) < -Bias_{Obs}(\hat{\tau}) = Bias_{Exp}(\hat{\tau}) < Bias_{Exp}(x)$. ■